Natural System Regional Simulation Model Sensitivity and Uncertainty Analysis

FINAL REPORT

Prepared for

South Florida Water Management District
3301 Gun Club Road
West Palm Beach, Florida 33406

By

INTERA Incorporated
1541 N. Dale Mabry Highway
Lutz, Florida 33548

January 10, 2007

Natural System Regional Simulation Model Sensitivity and Uncertainty Analysis

Srikanta Mishra [(1)], Neil Deeds [(1)], Patrick Tara [(2)],

Astrid Vreugdenhil [(1)], John Avis [(3)], and Nicola Calder [(3)]

INTERA Incorporated

Austin, TX [(1)]
Lutz, FL [(2)]
Ottawa, ON [(3)]

January 10, 2007

EXECUTIVE SUMMARY

This report details a demonstration of the applicability of sensitivity and uncertainty analysis techniques for the Natural System Regional Simulation Model (NSRSM). Uncertainty analysis is the probabilistic quantification of uncertainty in predicted outcomes, in this case hydrologic outcomes from the NSRSM. Analysis of uncertainty allows an increased confidence in the reliability of physical and economic estimates that are based on these outcomes.

The report begins with a description of the NSRSM, and includes the selection of key inputs and outputs for the study. The Regional Simulation Model (RSM), developed by the South Florida Water Management District, is a numerical model capable of simulating surface water and groundwater interactions in shallow water table environments. One specific application of the RSM is the Natural Systems Regional Simulation Model (NSRSM), which was designed to simulate the predevelopment hydrologic response. The model was constructed using a predevelopment land use condition and predevelopment topography.

Selection of uncertain input and output metrics is described next. NSRSM parameters for the current study were chosen from two land cover types, 511—Ridge and Slough Marsh as well as 712 – Mesic Pine Flatwood. The conveyance parameters were Manning's coefficient (n) and detention storage. The ET parameters were vegetation coefficient and extinction depth. The hydrologic parameter was storativity in the groundwater system. Finally, topography was varied among three separate cases: "low", "base", and "high" maps. A total of 11 input parameters were thus considered as uncertain. The output metrics were based on water stage and transect flow. For water stage, a location was chosen in each land cover type, with cell 25492 in the Ridge and Slough Marsh and 25087 in the Mesic Pine Flatwood. At these locations, the stage was averaged for a one-week time window within the representative year. Also, the average daily flow for the Tamiami transect (Ridge and Slough Marsh) and the T712_East transect (Mesic Pine Flatwood) for a one-week time window were selected as outputs. These selections resulted in four total output metrics.

Singular Value Decomposition (SVD)-based sensitivity analysis is discussed in the next report section. SVD-based sensitivity analysis provides more insight than a classical "one-off"

sensitivity analysis because it reveals parameter interdependencies. SVD involves the factorization of the sensitivity matrix to create matrices which define linearly independent groups of parameters and outputs. A vector of singular values is also created by the decomposition. These singular values indicate the relative importance of each parameter group. The inclusion and importance of parameters in the linearly independent groups provide insight into both parameter interactions and synergies, as well as the local sensitivity of output metrics to the parameters. The findings regarding particular sensitivities are useful in providing context to the global sensitivity analysis subsequent to the uncertainty propagation described below.

The next section of the report describes the characterization of parameter uncertainty. Distributions are chosen for those parameters that are considered uncertain in the analysis. The 11 input variables were each prescribed a distribution based on standard approaches for distribution selection, including a review of available literature, model calibration data from analogous regions, known constraints, and expert judgment.

In the next section of the report, uncertainty propagation techniques and results are discussed, along with a description of the software used to help automate the analysis. mCalc, the software used in the analysis, facilitates the evaluation of uncertainty or confidence in model predictions. It is a generic uncertainty analysis tool which 'wraps' around process models, enabling the user to develop input parameter samples from the input parameter distributions, feed the sampled parameter fields into the process model and compile and display CDFs of output variables (i.e., performance measure) of interest. mCalc implements the two uncertainty propagation techniques used in the current work, Monte-Carlo simulation (MCS) and First-Order Second-Moment (FOSM) analysis.

Monte-Carlo simulation, the most commonly employed technique for implementing the probabilistic framework in engineering and scientific analyses, is a numerical method for solving problems by random sampling. Probabilistic modeling allows a full mapping of the uncertainty in model parameters (inputs) and future system states (scenarios), expressed as probability distributions, into the corresponding uncertainty in model predictions (output), which is also expressed in terms of a probability distribution. Uncertainty in the model outcome is quantified via multiple model calculations using parameter values and future states drawn randomly from

prescribed probability distributions. Cases of 100, 200, and 300 realizations were run with the NSRSM model, and the results of the 200 realization case were chosen for detailed analysis. The CDFs for all of the output metrics were plotted, with most showing typical behavior. One exception was the [Tamiami] metric, which showed a considerable positive skew.

The First-Order Second-Moment (FOSM) method is one of the methodologies for estimating uncertainty in model predictions in terms of mean and standard deviation only, rather than the full output distributions. The advantage of this approach is that it typically requires only (number of parameters) + 1 model simulations, as opposed to several hundred simulations for typical MCS. For this particular case, one-point derivatives were not sufficient, so considerably more model simulations were required for calculating multi-point derivatives. The FOSM analysis was carried out for all of the variables, with the exception of topography, since categorical variables are not amenable to derivative calculations. The FOSM results were compared to the MCS results. The means and standard deviations were similar, with the exception of the [Tamiami] metric, where the standard deviation was somewhat different, likely due to the skewness in the output distribution.

The next report section describes the uncertainty importance analysis that was completed using the 200 realization MCS results. The objective of uncertainty importance analysis is to quantify the contribution from individual input parameters to the uncertainty (the spread or variance) of model predictions and determine the corresponding importance ranking. Two importance analysis techniques were used, stepwise rank regression and classification tree analysis. In stepwise regression, a linear response surface is fit between the rank-transformed input and output variables and a sensitivity analysis is performed on this "surrogate" model. In general, the order by which variables are added to the regression model corresponds to their order of importance. The relative contribution to variance by individual parameters can be compared via the standardized regression coefficient. The total goodness-of-fit for the regression model is characterized by the coefficient of determination (R^2). In general, the results of the regression analysis showed that one or two variables contributed predominantly to the variance in an output metric. The vegetation coefficient, Manning's *n*, and to a lesser extent, topography, were the most important contributors to variance.

Classification tree analysis is typically used to provide insight into what variable or variables are most important in determining the extremes of output data. A binary decision tree is at the heart of classification tree analysis. The decision tree is generated by recursively finding the variable splits that best separate the output into groups where a single category dominates. The importance of the variables is demonstrated by their order of split, with the variables at the top of the classification tree (the first variables split) considered more important than the variables involved in later splits. For this analysis, the upper and lower quartiles of the output data were the selected categories. The results of the classification tree analysis were similar to those of the regression analysis, with only one or two variables typically dominating the separation of the categories.

Overall, this work demonstrates the utility of several sensitivity and uncertainty analysis techniques for application to the NSRSM model. Uncertainty analysis provides a probabilistic description of the output metrics (i.e., range of possible outcomes and the likelihood of each outcome), given the current level of knowledge (or lack thereof) about the input parameters. This facilitates quantitative risk-based decision-making on the part of management. Sensitivity analyses help identify key input parameters that have the largest impact on uncertainty for the chosen output metrics. This allows resources to be focused on improving the understanding of the key processes and parameters. The results of the various approaches for sensitivity analysis were quite consistent, lending confidence to the conclusions. It is important to remember that while the techniques applied in this work are quite general, the specific quantitative results of the current study are valid only for the selected metrics and are also conditional on the assumptions about uncertainty characterization for model inputs. For future work, a more comprehensive characterization of input uncertainty is desirable, along with the selection of output metrics that can be directly related management decisions.

TABLE OF CONTENTS

TABLE OF CONTENTS (CONTINUED)

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (CONTINUED)

1.0 INTRODUCTION AND SCOPE

1.1 Natural System Regional Simulation Model

The Regional Simulation Model (RSM), developed by the South Florida Water Management District (the DISTRICT), is a numerical model capable of simulating surface water and groundwater interactions in shallow water table environments. The RSM was developed with a comprehensive hydraulic component capable of simulating the numerous and different types of man-made structures and canals in south Florida. The hydraulic component must be capable of responding to preset rules and operations as well as to extreme weather patterns (wet/dry) that affect competing urban, environmental and agricultural demands. The RSM is developed on a sound conceptual and mathematical framework, simulating a wide range of hydrologic conditions. The RSM has been developed principally for application in South Florida, and accounts for interactions among surface water and groundwater hydrology, structure and canal hydraulics, and management of these hydraulic components.

The RSM simulates and integrates the coupled movement and distribution of groundwater, surface water, man-made structures and network canals in south Florida. One specific application of the RSM is the Natural Systems Regional Simulation Model (NSRSM), which was designed to simulate the predevelopment hydrologic response. The model was constructed using a predevelopment land use conditions and predevelopment topography. All present day canals and structures were not represented in the NSRSM.

The NSRSM is intended to be a management tool to guide the restoration of the natural environment to predevelopment conditions. NSRSM applications are expected to produce an estimate of what the hydrologic state of the groundwater and surface water bodies would have been in the absence of man-made structures and network canals.

1.2 Need for Uncertainty Analysis

Uncertainty often refers to the error between model predictions and field observations. The fact that there is always doubt about the observed data as well as the model output accuracy

reduces confidence relative to the outcome of a particular hydrologic process. The quantification of the probabilistic characteristics of such errors is often called uncertainty analysis. There is a lot of natural variability and uncertainty in any model input, model parameters, model structure, and model output that contribute to the overall model uncertainty. In general there are three broad uncertainty categories.

1) Uncertainty due to our inability to fully understand the natural variability of input process to the model at a scale smaller than the gauging scale. Examples of these uncertainties are:
    a. Spatial variability such as rainfall, potential evapotranspiration (PET), and topography
    b. Temporal variability such as inflow and tidal boundary conditions
2) Uncertainty due to measurement errors and/or data modeling. This covers:
    a. All field measurements and published data, which are used to set up the model (e.g., input data and parameters)
    b. All field measurements used to calibrate the model (e.g., output).
    c. Data estimation using pre-processing and modeling tools
3) Uncertainty due to conceptual and implementation errors. This includes:
    a. Error in specifying boundary conditions such as inflow and tidal boundaries and initial conditions such as stage
    b. Model structural and numerical errors
    c. Model parameter errors due to parameter modeling errors and/or calibration imperfection
    d. Model inability to resolve variability smaller than the designated time step and mesh cell size
    e. Model linkage to other tools such as HPM that process stresses (e.g. crop local demand and runoff applied to the model)

Of interest is to identify, isolate, and quantify those sources of uncertainties with significant and unique contribution to the overall uncertainty associated with model output.

The uncertainty assessment of NSRSM is mainly aimed at the computation of the probability density function of key model output and performance measures which can help the water manager infer the following:

1. In its simple format, a mean and a standard deviation of a given output, performance measure or index. This simplified uncertainty metric is rarely sufficient for a complete characterization of uncertainty.
2. Model output in terms of a range rather than a single value. This describes the system performance as a range of potential outputs, classes of likely events, or probability density function.
3. Provides a level of confidence that a certain output is within an acceptable performance indicators.
4. Provides probability that a certain output exceeds a specific target value.

Previously, the DISTRICT has carried out a limited uncertainty and sensitivity analysis for the NSM, i.e., the predecessor of NSRSM. In this analysis, a sensitivity matrix was constructed to summarize the model response at each observation point to each individual performance parameter. A matrix factorization technique (Single Value Decomposition) was applied to the sensitivity matrix to isolate parameter groups with distinct parameter contributions. Model output uncertainty was then computed as a function of model parameter uncertainty using first-order error analysis.

The DISTRICT has had several workshops on uncertainty and sensitivity analysis. A recent peer review of the DISTRICT current modeling tool (i.e., SFWMM) has provided the following specific recommendations to be adapted for future improvement in this area for both the SFWMM and the NSM. The recommendations call for:

1) State clear and realistic objectives of sensitivity and uncertainty Analysis for SFWMM and NSM.
2) Conduct global sensitivity analysis to identify the parameters and variables with significant contribution to model output.
3) Incorporate uncertainty measures for both model parameters and model performance indices during model initial calibration.

4) Use uncertainty measures that are effective, quantitative, mathematically consistent and acceptable.
5) Produce a document that clearly provides answers to the uncertainty associated with model output, stability of model parameter estimates, the impact of parameter uncertainty on model output, and the relative contribution of different parameters to the overall model uncertainty.

1.3 Scope of Study

The objective of the current project is to demonstrate the application of uncertainty and sensitivity analysis techniques to NSRSM by computing the uncertainty associated with model output(s) as a function of prescribed uncertainty in selected model inputs and model parameters. The emphasis will be on a few key model inputs and parameters to which the NSRSM performance measures of interest are most sensitive to. To this end, a systematic framework for sensitivity and uncertainty analysis will be employed which comprises of the following steps:

- Selection of a limited set of key inputs and outputs based on previous modeling studies and expert opinion.
- Application of formal local sensitivity analysis (via Singular Value Decomposition of the input-output sensitivity matrix) to identify important uncertain inputs.
- Assignment of probability distributions to characterize uncertainty in selected model inputs and their correlation structure (based on the best available data).
- Application of uncertainty propagation techniques to determine the uncertainty in model output(s) as a function of the uncertainty in model inputs.
- Application of global sensitivity (uncertainty importance) analysis techniques to identify those model inputs that are key contributors to the overall uncertainty in model output(s). This results in an importance ranking that is dependent on both input uncertainty and input-output sensitivity, whereas the importance ranking based on SVD factorization is only dependent on input-output sensitivity.

2.0 MODEL DESCRIPTION

2.1 Description of RSM and NSRSM

The DISTRICT completed an initial numerical model implementation for South Florida. The numerical code currently being implemented has been under development for approximately ten years. This model code is titled the Regional Simulation Model (RSM) and is currently composed of two principal components that include the Hydrologic Simulation Engine (HSE) and the Management Simulation Engine (MSE). The HSE and MSE are coupled within the RSM C++ object-oriented code and do not exist as separate models (Figure 2-1). User input dictates if MSE components are used in conjunction with an HSE simulation. At this time, the RSM model is running only on the Red Hat Linux 9.0 platform, while pre- and post-processing codes run on both Linux and Windows platforms.

The RSM is a regional model that will be used to predict the hydrologic responses to planning and operational scenarios while considering competing water management priorities and issues. The HSE simulates the coupled movement and distribution of groundwater and surface water throughout the model domain. The MSE provides methods that can simulate operational decisions and/or alternative management decisions for the regional water distribution system. This model represents the next generation of integrated water management modeling and provides the ability to simulate the complexity of the South Florida hydrologic system and is necessary to support decision-making processes well into the future.

The RSM regional implementation to pre-drainage system (i.e., the natural system) is called the Natural System Regional Simulation Model (NSRSM). The NSRSM is a tool which mimics natural hydrologic conditions in south Florida. The primary goal of the NSRSM is to simulate pre-drainage hydrology within the estimated range of performance documented in the best available information sources. Its predecessor, the South Florida Water Management Model (aka the 2x2) based NSM, which was used extensively during the Central & Southern Florida Project Restudy (C&SFR), Comprehensive Everglades Restoration Plan (CERP), and several water supply planning efforts, to provide insight in evaluating alternatives for future restoration initiatives. The NSRSM will have the advantage of improved data sets and refined parameter

input resulting in simulations more closely representing natural system hydrology prior to modern drainage activities. Its performance will be evaluated for an extended period of record simulation that will represent system conditions starting mid 1800s to present time.

The NSRSM is a distributed model with spatial coverage that includes Lower Kissimmee Basin, Lake Okeechobee, Greater Everglades, Big Cypress, and Lower East Coast. The NSRSM is the implementation of the HSE used for the new generation management model as applied to the natural system after updating all input data and model parameters. The HSE is based on a weighted implicit finite volume approximation of the vertically averaged overland flow equations (i.e., Saint Venant equations). These equations consist of a continuity equation and momentum equations. For more information regarding the HSE methodologies the reader is referred to the RSM theory manual (http://www.sfwmd.gov/site/index.php?id=681)

Input data to the model include dynamic data such as historical rainfall, estimated evapotranspiration, and boundary conditions as well as static data such as topography, land cover, and aquifer thickness. Input parameters include groundwater parameters such as hydraulic conductivity, storage coefficient, seepage parameters, and surface water parameters such as Manning's coefficient, and Hydrologic Process Module (HPM) parameters such as wetland system's parameters, unsaturated soil parameters, and river network parameters.

The predevelopment land use was estimated from historical data collected from the General Land Office (GLO) surveys (land use data shown in Figure 2-2). The topography for the NSRSM was estimated with modern topography data in most places except for regions where significant subsidence had occurred, the topography in these areas utilized the GLO surveys.

The NSRSM was conceptualized with a very refined mesh. The model represents the predevelopment surface water hydrology and the surficial aquifer system. The lower boundary to (and from) the deep aquifers is assumed no flow. The model simulates the hydrologic (rain and ET) conditions from 1/1/1965 to 12/31/2000. There is a network of rivers simulating only the natural stream channels.

NSRSM simulates water level (stage) as the primary variable and estimates flows at key locations. Model Performance measures are hydroperiod, water depth, seasonal amplitude, flow

magnitude, and flow directions. There are two main simulation scenarios. A baseline condition run, which utilizes the SFRSM historical rainfall (i.e., 1965 through 2000), and an extended period of record run, which utilizes a longer period of records (i.e., 1895 through 2005) for the purpose of estimating extreme wet, dry, and average hydrologic conditions. Performance measures are compared against reference ranges for performance evaluation from best available sources of historical observations and measured data.

2.2 Selection of Key Inputs and Outputs

2.2.1 Land Cover Types

In order to keep the analyses tractable, the sensitivity and uncertainty analyses will be restricted to two unique land cover conditions. These two land cover types, *511--Ridge and Slough Marsh* and *712--Mesic Pine Flatwood*, rank 1 and 2 in terms of covered area in the NSRSM data set. Figure 2-3 shows the location of these two land cover types. Because these cover types comprise such a large area in the model, the variation of parameters within these areas should provide a relatively well-integrated impact on many of the potential output metrics.

2.2.2 Evaluation of Input Parameters

Based on our evaluation of project documents supplied by the DISTRICT, we developed a preliminary list of input parameters to be treated as uncertain. The input parameters were chosen from several water budget categories, e.g., conveyance, evapotranspiration (ET), and groundwater. These major categories provide a good cross section of the hydrologic processes captured in the HSE simulation.

a) Conveyance Parameters

Conveyance or overland flow is a significant surface water process in the overall water budget of a basin. The amount of water that is conveyed in overland flow is controlled by several hydrologic parameters and follows Equation (2-1):

$$Q = \frac{L}{n} d^{\frac{5}{3}} \sqrt{S} \qquad (2\text{-}1)$$

where

L = length of the flow face perpendicular to the flow direction,

n = Manning's coefficient,

d = water depth, and

S = water surface slope.

Excess precipitation or the rainfall that exceeds the infiltration capacity of the soil is available to runoff the basin via overland flow. The runoff processes are controlled by topography through the slope and the hydraulic length. The parameters selected to be treated as uncertain are the Manning's n and the detention storage.

Conveyance: Manning's n – Manning's n is related to the friction in the overland flow process. The HSE defines Manning's n using the following equation:

$$n = Ad^{B} \tag{2-2}$$

where

d = water depth, and

A, B = empirical constants.

In the case of the NSRSM, A is defined for each land cover type while B remains zero for all land cover types. In this case the depth variation of the n is ignored, so n is simply equal to A.

Conveyance: Detention Storage – Detention storage is defined as the minimum depth of surface ponding required in order to produce overland flow. The detention storage accounts for the micro-topography not represented by the topography defined by the scale of the cells. The detention storage basically acts as a switch. When the ponding is less than the detention storage then the overland flow is set to zero. When the ponded water exceeds the detention storage overland flow occurs following the equation above. The detention storage parameter was selected to be handled as uncertain.

b) ET Parameters

Evapotranspiration is the largest water budget term in the hydrologic cycle. Much of the water that enters the hydrologic cycle in the form of precipitation leaves in the form of evaporation. The Hydrologic Process Module or HPM dataset contained within the NSRSM includes the "layer1nsm" and "unsat" modules. The HPM contains the ET parameters associated

to the land use condition in each cell. There are several parameters associated with the "layer1nsm" and "unsat" HPM's. A graphic of the layer1nsm is extracted and shown below in Figure 2-4. Two parameters, 'kveg' and 'xd', were selected to be treated as uncertain. These parameters, as described below, are used by both "layer1nsm" and "unsat" modules.

ET: Vegetation Crop coefficient – The crop coefficient, kveg, represents a factor used to define the plants maximum capability to transpire water. The kveg parameter was selected to be handled as uncertain. The coefficient is not directly measurable and can only be determined through calibration.

ET: Extinction depth – The extinction depth, xd, is defined as the water table depth at which ET ceases to remove water from the water table and vadose zone. The ET crop correction factor linearly approaches zero starting from the root depth at which point the ET factor is defined as kveg. In the HSE formulation the extinction depth accounts for the dwindling number of roots at depth by further reducing the ET factor and thus the ET rate for the cell. Note that this parameter is a calibration parameter. There is no direct measurement of the extinction depth.

c) Hydrogeologic Parameters

The NSRSM uses a one layer model to represent the groundwater system. This one layer has a tidal flow boundary on the east coast to represent the flow to tide and a no flow boundary on the bottom face. The hydraulic conductivity of the surficial is represented in the "hyc-1a.dat". The storage parameter in the NSRSM is defined using the land cover index and the sv converter module. The "sv45.xml" is called from the main xml to perform this operation. For most land cover conditions the storage is considered as a constant. For the Ridge and Slough land use (index 511) the storage is defined with a look up table. In this project we will consider the storage parameter as uncertain.

2.2.3 A Note on Topography

The uncertainty in topography and how it affects the simulation results is a high priority for the DISTRICT. Unfortunately the DISTRICT cannot at this time provide us with many realizations of topography that would be required to properly capture its uncertainty and determine the corresponding uncertainty ranges in the model output. There is an on going

project that will be able to produce the many realizations but the timing of this project will not allow this project to use those realizations. However the DISTRICT has agreed to supply INTERA with two extreme scenarios around the reference topography that permit bracketing the range of topographic uncertainty and its use in uncertainty propagation studies.

Figures 2-5 and 2-6 depict the difference between the low and the high topography delivered to INTERA. These figures show that low topography ranges from no change to 0.27 feet lower and the high topography ranges from no change to 0.56 feet higher that the original topography. Topography was not considered in the local sensitivity analysis, due to the difficulty in "perturbing" a topography based on a limited number of available scenarios. In other words, a local tangent approximation cannot be applied as topography has not been parameterized in terms of a continuous and differentiable variable. However, topography will be considered as part of the global sensitivity and uncertainty analyses using the three different topo maps supplied by the DISTRICT with each of the maps assigned a different probability (weight).

2.2.4 Input Parameter Summary

Tabulated below in Table 2-1 is some basic information about the data file, source and range of values in the current NSRSM model associated with each of the parameters selected in the preceding paragraphs. Note that the current model documentation reflects the deterministic nature of the NSRSM model, and as such emphasizes considerations of spatial variability rather than lack-of-knowledge uncertainty. Further information about uncertainty ranges will be presented in Section 4 based on interactions with the DISTRICT staff, and/or evaluation of literature values.

Table 2-1 Summary of selected parameter sources and ranges.

Selected Parameter	Data File	Data Ranges in NSRSM	Data Source (author)
Topography	nsmtopo_v4.1.dat	-2.9 - 168.8	Appendix_Topo.doc (Winifred Said)
Conveyance – 'a' parameter	conveyance45.xml	.1 - .4 mostly .3	
Conveyance – detention storage	conveyance45.xml	.1	
ET – 'kveg'	hpm45.xml	-.1 – 1.0 [1]	
ET – Extinction Depth	hpm45.xml	3 – 10	
Hydrogeology – 'specific yield'	sv45.xml	.2 – 1.0 mostly .2	Appendix_Hydro.doc (Emily Richardson)

[1]Value -0.1 is used when kveg (ET) has seasonal variation

The selected parameters from these processes and their nominal values are listed in Table 2-2, along with the abbreviations by which they are referenced in the sensitivity analysis and uncertainty analysis results sections.

Table 2-2 Preliminary list of uncertain input parameters.

Parameter	Original Value	Description	Abbreviation
Alpha-712	0.3	Conveyance – *a* parameter	alpha712
Alpha-511	0.325	Conveyance – *a* parameter	alph511
Detent-712	0.1	Conveyance – detention storage	detent712
Detent-511	0.1	Conveyance – detention storage	detent511
Xd-712	10	ET – extinction depth	xd712
Xd-511	3	ET – extinction depth	xd51
Kveg-712	0.74	ET – kveg	kveg712
Kveg-511	0.74	ET – kveg	kveg511
Storativity-712	0.2	Hydrogeology – specific yield	sv712
Storativity-511	0.8	Hydrogeology – specific yield	sv511

2.2.5 Output Parameters

In a typical uncertainty analysis, the selection of output metrics is often guided by the metrics that define the targets used in model calibration. This consistency allows an analysis to yield insights about acceptable parameters ranges for key uncertain inputs. To this end, we initially chose output metrics that were similar to the soft calibration targets developed by the NSRSM team (e.g., daily maximum and minimum stage during a representative year). However, when we attempted to calculate numerical derivatives for the SVD analysis, we found that the chosen metrics did not respond smoothly to perturbation in the input parameters – possibly because of the non-continuous and non-differentiable nature of extreme values used as the metrics. This also affected the computation of derivatives for uncertainty propagation using the first-order second moment method. Because the metrics for the uncertainty analysis should be consistent with those used in the SVD analysis for a proper comparison, it was necessary to find alternative metrics that would respond more smoothly to perturbation. The exploratory work that yielded the final metrics is described in a separate document. The model outputs finally chosen as metrics for the uncertainty analysis are two cell stages, and two transect flows. For both the stages and transect flows, the daily result was averaged over a selected week during the representative period.

The stages at locations 25492 (land cover 511) and 25087 (land cover 712) were chosen as output variables of interest. Figure 2-3 shows their locations. Figure 2-7 shows the stage hydrographs at these locations for the representative water year ranging from May 92 to April 93, and the time "windows" over which the stages were averaged to yield the output metrics. The water year was defined by the DISTRICT and is representative of an average or normal rainfall condition.

In addition to the stages, a transect was chosen from each of the land cover types. For land cover 511, the flow from the Tamiami transect was selected. For land cover 712, no transect was immediately available, so one was created and approved by the DISTRICT. This transect will be called T712_East. The location of both transects is shown in Figure 2-3. Figure 2-8 shows the streamflow hydrographs for these transects for the representative water year ranging from May 92 to April 93, and the time "windows" over which the stages were averaged to yield the output metrics.

With two stage output metrics and two transect flow metrics, we had a total of four output metrics. Table 2-3, below, shows a summary of these outputs and the names by which they are referenced in the results section.

Table 2-3 List of output metrics and names.

Metric Type	Location	Abbreviation
Stage	25492, land cover 511	25492stage
Stage	25087, land cover 712	25087stage
Transect Flow	land cover 511	Tamiami
Transect Flow	land cover 712	T712_East

3.0 SVD-BASED SENSITIVITY ANALYSIS

3.1 Background

Classic local sensitivity analysis involves perturbing input parameters, typically one at a time, and observing the change in prescribed output metrics. By comparing the relative change in the output metrics to each parameter perturbation, one can gain insight into the relative sensitivity of output metrics to the input parameters. The disadvantage of this classical "one-off" sensitivity analysis approach is that it fails to take into consideration the effect of parameter interdependencies. Model input parameters often exhibit strong correlation and/or synergistic effects on the model output, which should always be accounted for during interpretation of input-output sensitivities.

Insight into parameter interdependence can be gleaned by further analysis of the input-output sensitivity data. One simple and effective method is singular value decomposition (SVD) of the sensitivity matrix. SVD involves the factorization of a rectangular matrix (in this case, the sensitivity matrix). It is a powerful technique that has been predominantly used in the field of hydrologic modeling for parameter reduction in the optimization process and also to improve convergence (e.g. Doherty, 2004). The added benefit of describing parameter resolution and interdependence has also been emphasized in some cases (Lal, 1995; Trimble, 1995). In the current work, we will apply the technique to enhance the sensitivity analysis of the NSRSM.

3.2 Theoretical Considerations

We start with a sensitivity matrix, $A_{m\,x\,n}$, where the elements are

$$\alpha_{ij} = \frac{\partial y_j}{\partial x_i} \qquad i = 1,2,\ldots,n \qquad j = 1,2,\ldots,m \tag{3-1}$$

where

α_{ij} = the sensitivity of the j^{th} simulated output metric to the i^{th} parameter

y_j = the j^{th} simulated output

x_i = the i^{th} parameter

m = number of observations

n = number of parameters

This sensitivity matrix is often referred to as the Jacobian. The matrix A can be decomposed into three matrices V, S, and U such that

$$A = U S V^T \tag{3-2}$$

U and V are m x m and n x n matrices, respectively, and S is a diagonal matrix of singular values of A. V^T gives the coefficients of linear combinations of the original parameters that give rise to new, independent parameter groups. U gives the coefficients of linear combinations of the observation groups. The parameter groups and observation groups are related by the diagonal matrix S. The relative magnitude of the singular values in S indicates the relative importance of each of the parameter groups.

Typically a cutoff level is established for the smallest singular value in S for controlling data errors. For this study, we will use the rule of $s_{min}/s_{max} < 0.001$ (Trimble, 1995). Values of s that are less than s_{min} are set to zero, and the number of columns in the U and V^T matrices are correspondingly reduced.

The resolution matrix, R, where

$$R = V V^T \tag{3-3}$$

gives insight regarding parameter resolution (or parameter independence). Also, the correlation matrix can be determined from the SVD via the covariance matrix Y:

$$\rho_{ij} = \frac{Y_{ij}^2}{Y_{ii} Y_{jj}} \tag{3-4}$$

where ρ_{ij} is element i,j of the correlation matrix.

The singular values, U and V^T, the resolution matrix, and the correlation matrix are the primary sources of information used to construct groups of parameters, understand their interdependence, and analyze their sensitivity.

3.3 Results

The SVD sensitivity analysis was completed on the parameters described in the previous section. Recall that there are 10 input parameters (since the pointer variable for topographic uncertainty is kept unchanged), and 4 output metrics. Producing the sensitivity matrix requires calculating the derivatives expressed in Equation 3-1. In a typical analysis, this requires running the forward model N+1 times, where N is the number of input parameters. In each run, one of the parameters was perturbed 1% from its original value, and numerical derivatives are calculated. However, as we indicated in Section 2.2.5, output metric responses to small perturbations in inputs were not consistently smooth or monotonic, which could result in considerable error in calculating a one-point numerical derivative. Therefore, for this work we calculated multipoint derivatives, using 14 randomly-spaced points inside a span -5% to 5% from the base parameter value. A linear regression was performed on the points inside the span, and the derivative was calculated as the slope of the regression line. This approach is obviously less computationally efficient than N+1 for the single-point derivatives, but was necessary to ensure accuracy for this case. Each element of the sensitivity matrix was divided by the standard deviation of the output metric (based on the set of daily values in the 1992-93 average year) to remove unit dependencies (Lal, 1995).

3.3.1 Sensitivity Matrix

A visualization of the sensitivity matrix is shown in Figure 3-1. The size of the bubble in the plot is proportional to the magnitude of the element in the sensitivity matrix. The results are reasonably self-consistent; that is, the two metrics in land cover 511 are sensitive primarily to [alpha511] and [kveg511] and the two metrics in land cover 712 are sensitive primarily to [alpha712] and [kveg712]. So the Manning's *n* and vegetation coefficients are the most influential parameters for this set of metrics. The [25492] metric shows a similar relative response to the [Tamiami] metric, while the [T712_East] metric has a larger relative response than the [25087] metric. However, keep in mind that this response is somewhat dependent on the choice of the "time window" over which the metric was averaged, and the standard deviation of the output over the representative year. Because of the difficulty in normalizing metrics representing different physical processes and units (i.e. stage versus transect flow) at different

times, the strength of the SVD technique lies not in identifying the metric that is most sensitive to a particular parameter, but rather in identifying the most influential parameters *and their potential interactions* with respect to a particular metric.

3.3.2 Singular Values

Figure 3-2 shows the singular values resulting from the decomposition. The figure indicates that the first parameter group is the most influential group, followed by significantly lower singular values for the remaining groups. Using our defined cutoff of $s_{min}/s_{max} < 0.001$, we will keep all 4 singular values and will therefore be analyzing all 4 parameter groups.

3.3.3 Output Metric Groups

Figure 3-3 shows the elements of the U matrix which provide the linear coefficients making up the output metric groups. Coefficients of higher magnitude indicate more important metrics for a particular group. The trend in Figure 3-3 is quite clear, and corresponds to the observations made from Figure 3-1. Recall that the order of the groups corresponds to the order of the singular values; therefore, they are numbered and displayed in order of decreasing overall sensitivity. The first group, O1, is dominated by the output metrics [T712_East], which had the largest relative response in Figure 3-2. Groups O2 and Group O3 are dominated by [Tamiami] and [25492], respectively. These metrics showed responses of similar magnitude in Figure 3-2. Finally, Group O4 is dominated by [25087] which displayed the smallest relative response in Figure 3-2. These figures indicate that the output groups are all mostly defined by a single metric, so the matching input groups (discussed in the following section) will correspond predominantly to a single metric.

3.3.4 Input Parameter Groups

Figure 3-4 shows the elements of the V_T matrix which provide the coefficients for the linear combinations of parameters that make up the input parameter groups. Coefficients of higher magnitude indicate more influential parameters for a particular group. Note that these groups are ordered to correspond with the singular values and therefore with the output groups. The two most influential parameters in Group I1 are [alpha712] and [kveg712], with a smaller contribution by [detent712]. Recall that the corresponding Group O1 was dominated by metric

[T712_East], so this result is similar to the general trend shown in the sensitivity matrix in Figure 3.1. The dominant parameters in Groups I2 and I3 are [alpha511] and [kveg511] which are consistent with previous observations for metrics [Tamiami] and [25492stage] corresponding to Groups O2 and O3. Finally, the dominant parameters in Group I4 are [alpha712] and [kveg712]. This result is expected given that metric [25087] dominates Group O4.

These results from the elements of the V^T matrix show that the Manning's n and vegetation coefficients work in tandem in influencing the output metrics that reside in respective land cover types.

3.3.5 Resolution Matrix

Figure 3-5 shows a visualization of the resolution matrix, which indicates how well resolved each of the parameters is with respect to the output metrics. A well-resolved parameter is one that can be independently determined without significant influence from other parameters. If all of the parameters were independent, the matrix would be an identity matrix (i.e., unity along the diagonal). We can see from Figure 3-5 that [alpha511] is the best resolved parameter, likely due to its significant importance in both Groups I2 and I3. The other well-resolved parameters are also important in two groups, with [alpha712] and [kveg712] in Groups I1 and I4 and [kveg511] in Groups I2 and I3. If we look at a parameter that is poorly resolved, such as [xd511], we note that it barely registers in any of the parameter groups. Remember that this result does not mean that the parameter is not influential (although the sensitivity matrix indicates it is of marginal influence), but rather that its importance is not independent of other parameters in the analysis.

3.3.6 Correlation Matrix

Figure 3-6 shows the correlation matrix. This matrix indicates which parameters have a similar direction of influence on the output metrics (not necessary a similar magnitude of influence). In many cases, the correlation matrix will support the resolution matrix, with well-resolved parameters showing the least correlation to other parameters. This trend is evident here with the [alpha511] and [kveg511] parameters, which show only slight correlation to other parameters. However, the other well-resolved parameters [alpha712] and [kveg712] do show

significant correlation to some of the other parameters. As noted earlier, although the direction of influence may be similar between two parameters, the significant difference in magnitude of influence may result in one parameter being well-resolved and the other not.

3.4 Discussion

One of the potential benefits of the SVD-based sensitivity analysis is to identify those parameters that exhibit the highest input-output sensitivity. Such parameters can then be utilized during model calibration studies and make the parameter estimation process parsimonious and stable. SVD-based parameter screening can also be used to restrict the number of parameters that are selected for probabilistic modeling studies. The parameter screening aspect was not relevant in the present study because of the limited number of uncertain inputs considered. It should be kept in mind, though, that the SVD-based sensitivity analyses reflect results that are representative of conditions at or near the reference state, and may not be accurate away from the reference point if the input-output relationship is nonlinear and/or non-monotonic.

In a later section, we will present results of uncertainty importance or global sensitivity analyses that provide a complementary suite of results compared to those from the SVD-based sensitivity analyses. As noted earlier, the global sensitivity analyses produce importance rankings that are dependent on both input uncertainty and input-output sensitivity, whereas the importance rankings based on SVD factorization are only dependent on input-output sensitivity. The global sensitivity analysis results will be generated from the sampling-based uncertainty analyses, and will thus reflect input-output sensitivities over the entire range of parameter variations. The other advantage of the global sensitivity analysis is the consideration of synergistic effects between parameters, whereas the sensitivities used in the SVD-based analysis are derived from one-parameter at a time perturbations.

4.0 CHARACTERIZATION OF PARAMETER UNCERTAINTY

4.1 Background

This section describes the distributions assigned to the key uncertain inputs identified in Section 2.2. Recall that the uncertainty analyses will be isolated to two unique land cover conditions (511, Ridge and Slough Marsh and 712, Mesic Pine Flatwoods). There are five uncertain parameters corresponding to each land cover condition, viz., Manning's *n*, detention storage, vegetation crop coefficient, extinction depth and storage coefficient. In addition, there is one indicator variable representing topographic uncertainty.

4.2 Conveyance – Manning's *n*

The distribution selected to represent the Manning's *n* for the Ridge and Slough Marsh and the Mesic Pine Flatwood is shown in the tables below. The original model values for Manning's *n* are shown in the parenthesis. The unique land cover of the Ridge and Slough Marsh makes a thorough literature review difficult. The "Predevelopment Vegetation Communities of Southern Florida" states that the Ridge and Slough community is an Everglades specific community. Since this environment is only found in the Everglades widely accepted literature values are difficult to find. All the data will be very specific to South Florida and work performed in South Florida. In contrast the Pine flatwoods vegetation community is wider spread and therefore literature values from various authors are a little easier to obtain. Also the range of possible conditions in this land use makes defining a single Manning's value problematic. Under very wet conditions lower ranges of friction factors are plausible due to the deep water depth. Overland flow frictions factors are no longer valid to represent the wet conditions possible. During dry conditions in the marsh a higher Manning's *n* is plausible.

The possible range of values along with the corresponding cumulative probability level is shown in the distribution of uncertainty in the Manning's *n* (shown in the table below). A review of various literature ranges for the pine forest showed higher values are possible. The range in Manning's *n* is defined by the range found in the literature. The literature showed an affinity towards a value of 0.4 for the Pine Flatwood, therefore, the distribution reflects this as a

predominate probable outcome. Piece-wise linear CDFs based on these values are shown in Figure 4-1, and will be used as inputs for uncertainty analysis.

Ridge and Slough Marsh (.325)		**Mesic Pine Flatwood** (.3)	
Value	CDF	Value	CDF
.06	0	.3	0
.3	.15	.35	.10
.35	.95	.45	.90
.4	1	.6	1

4.3 Conveyance – Detention Storage

The detention storage is defined as the minimum depth of surface ponding required in order to produce overland flow. The detention storage accounts for the micro-topography not represented by the topography defined by the scale of the cells, and basically acts as a switch. When the ponding is less than the detention storage then the overland flow is set to zero. When the ponded water exceeds the detention storage overland flow occurs following the equation above. The detention storage parameter was selected to be handled as uncertain.

The distribution of the detention storage for the Ridge and Slough Marsh and the Mesic Pine Flatwoods is shown in the following tables. The distribution for the detention storage was defined as uniform over the range. Due to the large variability in topography present in the ridge and slough marsh the detention storage was defined with a larger range. The Mesic Pine Flatwood has more regular and smooth topography therefore a smaller range in the detention storage was used to define the uncertainty. Figure 4-2 shows these piece-wise linear CDFs.

Ridge and Slough Marsh (.1)		**Mesic Pine Flatwood** (.1)	
Value	CDF	Value	CDF
.1	0	.1	0
.6	1	.2	1

4.4 ET – Vegetation Crop Coefficient

The crop coefficient, or kveg as it is defined in the input data set, represents a factor used to define the plants maximum capability to transpire water. Note that this parameter is not directly measurable and can only be determined through calibration.

The distributions of the uncertain inputs for the two selected land covers are shown in the tables below. The ridge and slough vegetation is capable of transpiring more than the pine flatwood vegetation. Values of similar vegetation coefficients for other models show the pine flatwoods can be quite lower than that of wetland marshes. Figure 4-3 shows these piece-wise linear CDFs.

Ridge and Slough Marsh (.88)	
Value	CDF
.7	0
.8	.5
.9	1.0

Mesic Pine Flatwood (.84)	
Value	CDF
.4	0
.6	.40
.7	.90
.8	1

4.5 ET – Extinction Depth

The extinction depth is defined as the water table depth at which ET ceases to remove water from the water table and vadose zone. The ET crop correction factor linearly approaches zero starting from the root depth at which point the ET factor is defined as kveg. In the HSE formulation the extinction depth accounts for the dwindling number of roots at depth by further reducing the ET factor and thus the ET rate for the cell. Again this parameter is a calibration parameter. There is no direct measurement of the extinction depth.

The distributions for the extinction depth uncertainty are shown in the tables below. The predominant vegetation in the ridge slough marshes do not allow for an extensive root system while the pine flatwoods can have extensive roots. The original values for the extinction depth are 3.0 and 10.0 for the marsh and pine flatwoods respectively. These numbers are reasonable given other hydrologic modeling studies. Even though the numbers are reasonable there are uncertainties associated with their selection. For example the ridge and slough vegetation could

easily have roots as deep as 4 feet or may be only as deep as 2 feet. But we do know the roots of the ridge and slough vegetation will not be 10 feet deep and they have to be deeper than 6 inches. Both land form parameter distributions were defined with normal distributions. The means are set equal to the original model parameter while the ranges and standard deviations allow for an adequate description of the possible values. The standard deviations are set by first establishing a plausible range, (i.e., 2-4 for the marsh, and 8-12 for the pine flatwoods), and then equating this range to 6 standard deviations.

Ridge and Slough Marsh (3.0)	**Mesic Pine Flatwood** (10.0)
2-4 Normal Distribution 3.0 mean .33 standard dev.	8-12 Normal Distribution 10.0 mean 0.667 stand dev

4.6 Storage Coefficient

For most land cover conditions the storage coefficient is considered as a constant. For the Ridge and Slough land use (index 511) the storage coefficient is defined with a look up table. The uncertainty in this parameter for the selected land forms are shown in the tables below.

The mucky soils associated with the ridge and slough land forms allow for fairly high storage factors. The defined distribution does test the possibility of lower ranges in the storage coefficient. Figure 4-4 shows this piece-wise linear CDF. The original .2 value of the mesic pine flatwood is fairly consistent with other models in Florida. A plausible range of .1-.3 is assumed for this variable, from which the standard deviation is estimated by equating the range to 6 standard deviations.

Ridge and Slough Marsh (.8 with lookup)		**Mesic Pine Flatwood** (.2)
Value	CDF	.1-.3 Normal Distribution .2 mean .033 stand dev
.5	0	
.6	.25	
.7	.50	
.8	1.0	

4.7 Topography

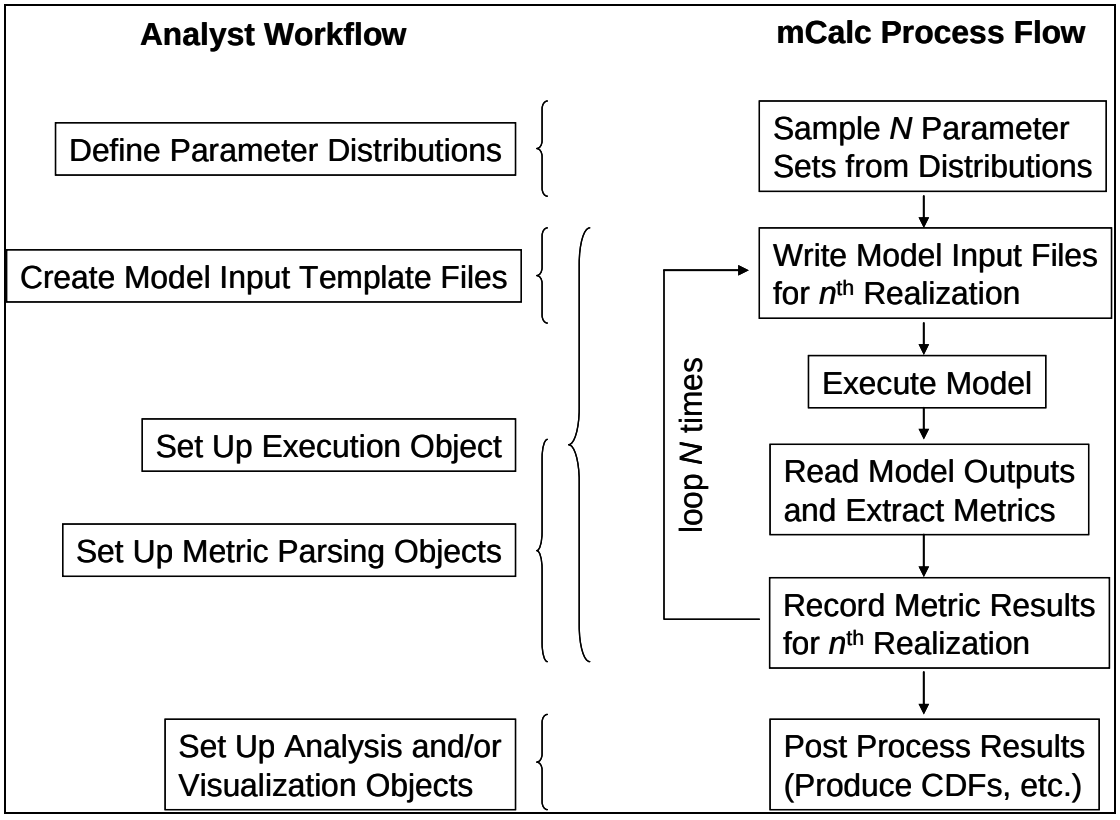
The DISTRICT has supplied INTERA with two extreme scenarios (maps) around the reference topography that permit bracketing the range of topographic uncertainty and its use in uncertainty propagation studies. Thus, the uncertainty in topography is characterized via 3 maps, a "low" topo map, a "reference" topo map, and a "high" topo map. An indicator variable called "toposelect" with values 1, 2 and 3 is used to represent the 3 topo maps. Based on the three-point discretization strategy of Keefer and Bodily (1983), we will apply a weight of 0.185 to the two end-member scenarios, and a weight of 0.63 to the reference scenario. In other words, the two end member scenarios will be invoked for 18.5% of the Monte Carlo simulations, and the reference scenario will be invoked for 63% of the Monte Carlo simulations in the uncertainty propagation studies. Thus, the uncertainty in topography is represented in terms of the discrete indicator variable [TOPOSELECT] and the probabilities as described above. This discrete distribution is shown in Figure 4-5.

It should be noted that although this simplistic representation of topographic uncertainty is adequate for demonstrating the application of NSRSM in a probabilistic framework, it results in the impacts of topographic uncertainty being restricted to a small spatial domain (as shown in Figures 2-5 and 2-6). Also, this discrete nature of uncertainty representation is really a coarse characterization of the true state of uncertainty. An alternative would be to generate multiple equi-probable geostatistical realizations that better characterize the spatial variability and correlation structure of topography in the model domain, as well as the uncertainty associated with this characterization. Such a representation could capture the uncertainty of topography in a more continuous and realistic manner.

5.0 UNCERTAINTY PROPAGATION

5.1 mCalc Software Description

The two uncertainty propagation methods used in the current work are Monte-Carlo simulation (MCS) and the First-Order Second-Moment method (FOSM) (Tung and Yen, 2005). Both techniques are integrated into the mCalc software, which is a powerful tool for examining model uncertainties. mCalc is a generic uncertainty analysis tool that 'wraps' around a process model (or a series of models). Originally developed as a validation tool in support of the Department of Energy's Yucca Mountain Project, mCalc provides a user-interface, a library of integrated sub-models, a sampling utility, an optimization utility, and many tools to analyze and visualize both field and simulated data. It was developed in accordance with strict software QA procedures, with each new component tested and verified in a manner appropriate to the component. The figure below shows a diagram of a typical analyst workflow and mCalc process flow for an MCS analysis. The figure is explained in the following paragraph.



In the left column of the figure, we see the analyst workflow. First, the analyst defines the parameter distributions in mCalc. mCalc has a wide variety of available distributions, and also allows the input of empirical CDFs. Next, the analyst creates model input template files, for those input files that contain uncertain parameters. A model input template file usually consists of the input file from the calibration run, with placeholders (that mCalc recognizes) at the locations of the uncertain parameter values. The metric parsing objects are flexible I/O objects that allow the reading of model output data and the subsequent extraction of relevant metrics. mCalc has many objects available that can read and process most types of output data. If mCalc does not have a suitable built-in object for reading the output data, simple scripts can be integrated into the execution loop for reformatting the output data. The execution object links all of the file creation, model execution, and metric processing objects together, so that they run in the correct sequence. Finally, the analysis and visualization objects are set up, that automate the creation of CDFs, horsetail plots, scatterplots and other visualization tools.

In the right column of the figure, we see the typical mCalc process flow. First, mCalc creates all of the samples from the defined parameter distributions. Either Latin Hypercube sampling, or Random (sometimes called Monte-Carlo) sampling can be used. Intra-parameter correlations can be specified for Latin Hypercube sampling. After the sampled dataset is created, mCalc starts the main processing loop. First, parameters for the first realization are drawn from the sampled dataset. The model input files are written based on these parameters and the model input template files. The model is then executed. Then the model output files are read and the relevant metrics are extracted and recorded. At this point, the parameters for the second realization are drawn, and the loop continues until all of the realizations have been run. Note that when mCalc is run on a cluster, this loop is the portion that is divided up and sent to the various nodes for parallel computation. Finally, when all of the realizations are finished, the analysis/visualization objects are used to automatically create the CDFs, scatterplots, and other relevant figures. mCalc also includes a suite of advanced statistical analysis tools for sensitivity (uncertainty importance) assessments, including correlation and partial correlation analysis, stepwise regression, and non-linear pattern detection.

mCalc can be used through its user-friendly GUI, or in batch mode from the command line. In batch mode, mCalc support parallel processing Linux or Windows MPI based clusters, minimizing the time required to run probabilistic simulations. Both 32-bit and 64-bit versions of mCalc are available, for both Windows and Linux platforms. For the current work, the mCalc configuration was created on Windows through the GUI and the simulations were run on a Linux cluster of seven nodes, each with Pentium D dual-core processors running at 3.2 GHz.

5.2 Monte Carlo Simulation

5.2.1 Background

Monte-Carlo simulation (MCS), the most commonly employed technique for implementing the probabilistic framework in engineering and scientific analyses, is a numerical method for solving problems by random sampling (Morgan and Henrion, 1990). Probabilistic modeling allows a full mapping of the uncertainty in model parameters (inputs) and future system states (scenarios), expressed as probability distributions, into the corresponding uncertainty in model predictions (output), which is also expressed in terms of a probability distribution. Uncertainty in model outcomes is quantified via multiple model simulations using parameter values and future states drawn randomly from prescribed probability distributions.

A Monte-Carlo analysis generally involves the following four steps:

1 Select imprecisely known model input parameters to be sampled
2 Construct probability distribution functions for each of these parameters
3 Generate a sample scenario by selecting a parameter value from each distribution
4 Calculate the model outcome for each sample scenario and aggregate results for all samples (equally likely parameter sets).

Step 1 has been previously described in Section 2.2. Step 2 is described in Section 4. mCalc allowed the automation of steps 3 and 4, as described above. mCalc was set up to run the NSRSM model in probabilistic mode by first creating the 11 variable objects corresponding to the inputs described in Section 2.2. Each of these objects was set up as an uncertain variable with the appropriate distribution parameters. The next step required creating template files for

the NSRSM input that would be modified during the probabilistic runs. Template input files are identical to the actual NSRSM input files, except that placeholder "tags" are added in locations were variable values occur. When mCalc encounters these tags in the template files, it replaces the tags in the created input file with the appropriate sampled values. Finally, mCalc objects were created to read the NSRSM output files and extract the metrics described in Section 2.2.

The Latin Hypercube sampling (LHS) option (McKay et al., 1979) was selected in mCalc for all runs, with no specified input-input correlations. Runs of 100, 200, and 300 realizations were completed, with 100 realizations taking approximately 25 hours of clock time.

5.2.2 Monte Carlo Simulation Results

In selecting the 200 realization case for analysis in this work, we compared a 100, 200, and 300 realization cases to examine stability (i.e., sensitivity of results to sample size). As examples, we will show a comparison for the [25492stage] metric. Figure 5-1 shows the CDF for the [25492 stage] metric for the 100, 200, and 300 realization runs. The horizontal grid lines mark the 5^{th}, 50^{th}, and 95^{th} percentiles, while the vertical lines mark the metric values that correspond to these percentiles. We can see from this figure that the 200 and 300 realization cases are smoother, and straighter through the middle section of the CDF than the 100 realization case. The similarity between the 200 and 300 realization cases indicates that an adequate number of samples have occurred to provide stable output statistics.

It should be noted that the CDF shown in Figure 5-1 reflects the uncertainty in model outcomes from realization to realization for the one-week time averaging window shown as vertical lines in Figure 5-2. In problems with time-dependent outputs, the selection of a representative "time slice" for presenting a CDF is often subjective. Depending on the objective of the modeling exercise and/or the choice of the decision variable(s), such time slices could be taken to correspond to the minimum, maximum, average or some other value over the simulation period. In this case, we have selected the time slice as a one-week averaging window at some temporal location that produces smooth derivatives so as to be consistent with the SVD analysis.

The remaining discussion will refer to the 200 realization case only. Figure 5-2 shows the "horsetail" plot for the stage at location 25492, over the period of the representative year. A

"horsetail" plot shows the uncertainty in a time-dependent outcome, typically along with several of the percentiles. In Figure 5-2, the 200 realizations are represented by the grey lines, the 5th and 95th percentiles are blue lines, the 50th percentile (or median) is represented by the red line, and the base run is represented by the green line. The base run represents the results from the NSRSM model provided by the DISTRICT, with variable values unchanged. Note that the median value corresponds closely to the base run, but is not exactly the same. Even if the base run parameter values were equal to the median values for the MCS run (which in this case, they were not), the median output would not necessarily result. The 95th percentile is very close to the maximum of the realizations where stage is high, while the 5th percentile lies along a much larger spread in results for the lower stage realizations. In general, the spread in the results appears to be largest in the winter months, in the middle of the period.

Figure 5-3 shows the horsetail plot for the stage at cell 25087. For much of the period, the spread is much smaller at this stage location, although the spread increases dramatically at two times, peaking in June of 1992 and again in January of 1993.

Figure 5-4 shows the horsetail plot for the Tamiami daily transect flow. In this figure we see that the maximum extremes can be significantly higher than the 95th percentile and the minimum extremes can be significantly lower than the 5th percentile, due to the occasional large fluctuation in flow with time. When the spread is smaller, the median tracks the base case run very closely.

Figure 5-5 shows the horsetail plot for the T712_East, which has the more sharply pronounced peaks than the Tamiami result. Compared to the scale of the largest peaks, the spread in the results does not appear very significant.

The horsetail plots show all of the model results for the representative year. To analyze the uncertainty in the actual metrics, we will discuss figures showing the CDFs for each of the metrics for the 200 realization case. As noted earlier, each CDF reflects the uncertainty in model outcomes from realization to realization for the one-week time averaging window shown as vertical lines in the horsetail plots. Figure 5-6 shows the CDF for the [25492stage] metric. The CDF is relatively symmetric, with a slightly longer tail near the lower bound. The mean at this location was 1.26 ft, while the standard deviation was 0.136 ft.

Similarly, Figure 5-7 shows the CDF for the [25087stage] metric. This CDF is quite symmetric, but has a longer tail in the lower percentiles, and lacks the familiar "S" shape, showing a relatively straight trend from about the 5th percentile onward. The mean at this location was 0.13 ft, and the standard deviation was 0.0465 ft. The "error bar" on the mean stage for the [25087stage] metric is therefore roughly 1/3 of that for the [25492stage] metric – which is consistent with the general spread of the uncertainty bands shown in the horsetail plots in Figure 5-2 and Figure 5-3. It should be noted that this comparison is at two different times.

Figure 5-8 shows the CDF for the [Tamiami] metric. This is not a typical CDF, starting with a large initial slope, and flattening suddenly near the 90th percentile. The mean for this metric is 2.70 x 10^{8} CFD, which is similar to the median, at 2.56 x 10^{8} CFD, in spite of the general asymmetry in the CDF. The standard deviation for this metric is 7.86 x 10^{7} CFD.

Similarly, Figure 5-9 shows the CDF for the mean [T712_East] metric. The mean for this metric is -8.25 x 10^{7} CFD and the standard deviation is 1.13 x 10^{7} CFD. The error bar on mean flow at this location is only about 15% of that for the [Tamiami] metric – which is consistent with the general spread of the uncertainty bands shown in the horsetail plots in Figure 5-4 and Figure 5-5. As before, it should be noted that this comparison is at two different times.

As noted earlier, the horsetail plots in Figures 5-2 through Figure 5-5 depict the uncertainty in model outcomes over the representative year, whereas the CDFs in Figure 5-6 through Figure 5-9 show the spread in model predictions at a particular point in time (in this case, a one-week averaging window). In order to examine the sensitivity of the shape of the CDF to exact placement of the one-week averaging window, we select two alternative time slices for the [Tamiami] metric – shown as [Alt1] and [Alt2] in Figure 5-5. The resulting CDFs are compared to the reference CDF in Figure 5-10. The reference CDF (denoted as [Metric]) has considerable right skew, because of the presence of outliers above the 95th percentile bound as shown in the horsetail plot in Figure 5-4. Conversely, the [Alt2] CDF has considerable left skew because of the presence of outliers below the 5th percentile bound. The [Alt1] CDF is more symmetric, because the outliers are in close proximity to the 5th/95th-percentile bounds. Thus, the degree of skewness seen in the CDF can be directly related to the behavior of the outliers in the horsetail plot.

5.3 First-Order Second Moment Analysis

5.3.1 Background

Often, it is sufficient to quantify the uncertainty in model predictions in terms of the mean (describing the central tendency of the prediction) and the variance (describing the spread around the mean) rather than the full distribution. The first-order second-moment method (FOSM) is one such methodology (Benjamin and Cornell, 1970; Dettinger and Wilson, 1981; Ang and Tang, 1984; Morgan and Henrion, 1990). Compared to MCS, the FOSM technique entails considerably less computational effort for problems with a small number of uncertain parameters, while providing results of comparable accuracy for linear and mildly nonlinear problems (Mishra and Parker, 1989; James and Oldenberg, 1997).

Consider an uncertain quantity, F, which depends on the parameter vector, $\mathbf{x}$. A first-order Taylor expansion around the mean point, $\hat{\mathbf{x}}$, gives:

$$F(\mathbf{x}) \cong F(\hat{\mathbf{x}}) + \sum_i \frac{\partial F}{\partial x_i}(x_i - \hat{x}_i) \tag{5-1}$$

with $\hat{\mathbf{x}}$ being the vector of mean values of the uncertain parameters, where the partial derivatives in Eq. 5-1 are also evaluated. Taking the expected value of both sides of this expression yields:

$$\mathbf{E}[F] \cong F(\hat{\mathbf{x}}) + \sum_i \frac{\partial F}{\partial x_i}\mathbf{E}[x_i - \hat{x}_i] \tag{5-2}$$

where $\mathbf{E}[\cdot]$ denotes the expectation operator. Assuming small and symmetrically distributed parameter perturbations around the mean values such that the expectation term in the RHS can be dropped and all higher-order terms neglected, we obtain:

$$\mathbf{E}[F] \cong F(\hat{\mathbf{x}}) \tag{5-3}$$

Thus, the first-order estimate of the expected value (mean) of the uncertain quantity, F, is obtained simply by using the mean (expected value) of each of the uncertain parameters to evaluate the model.

The variance of F is defined as:

$$V[F] = \sigma_F^2 = \mathbf{E}[(F - \mathbf{E}[F])^2] \tag{5-4}$$

which can be calculated by substituting Eqs. 5-1 and 5-3 in Eq. 5-4 as follows:

$$V(F) \cong \sum_i \sum_j \frac{\partial F}{\partial x_i}\frac{\partial F}{\partial x_j}\mathbf{E}[(x_i - \hat{x}_i)(x_j - \hat{x}_j)] \cong \sum_i \sum_j \frac{\partial F}{\partial x_i}\frac{\partial F}{\partial x_j} C[x_i x_j] \tag{5-5}$$

where the covariance, $C[x_i x_j] = r_{ij}\sigma[x_i]\sigma[x_j]$, is expressed in terms of the parameter correlation coefficients, r_{ij}, and the individual parameter standard deviations, σ. The variance of F is thus seen to depend on the variance-covariance relation of the input parameters, and its sensitivity to the uncertain inputs. For uncorrelated parameters, the expression for variance simplifies to:

$$V(F) \cong \sum_i \left(\frac{\partial F}{\partial x_i}\right)^2 V[x_i] \tag{5-6}$$

The first-order estimate of the mean given in Eq. 5-3 is a reasonable approximation so long as parameter variances are small and the function is only mildly nonlinear - which allows higher-order terms to be dropped. If these conditions are not met, then second-order terms need to be retained in the Taylor expansion of Eq. 5-1, leading to a correction to the mean which depends on the parameter covariance and mixed second partial derivatives (Benjamin and Cornell, 1970; Dettinger and Wilson, 1981).

5.3.2 FOSM Results

The FOSM method is automated in mCalc. However, as described in the SVD section, one-point derivatives did not produce accurate results. Because of this, the derivatives were calculated using the multipoint method described previously in Section 3.3. This makes the approach less attractive in terms of computational efficiency. Still, we completed the calculations so that the results from the FOSM method could be compared to the MCS results. All of the same uncertain inputs were used, with the exception of TOPOSELECT. TOPOSELECT is a categorical input variable, so its derivatives are undefined.

The results of the FOSM method are compared against those obtained using MCS in Table 5-1. In general, the estimates of mean and standard deviation from both methods are in good agreement for most of the output metrics. The only exception is the standard deviation for

the [Tamiami] metric, with the FOSM method estimate at less than 50% of the MCS value. This is likely due to the large skewness in the output distribution (Figure 5-8). The FOSM method requires a certain degree of symmetry to produce accurate results, which is one of the limitations of the method.

Table 5-1 Comparison of MCS and FOSM results.

	FOSM		MCS	
	Mean	Stdev	Mean	Stdev
25492stage	1.26	0.12	1.26	0.14
24087stage	0.13	0.054	0.13	0.047
Tamiami	2.55E+08	3.25E+07	2.70E+08	7.86E+07
T712_East	-8.49E+07	9.80E+06	-8.25E+07	1.13E+07

Recall that additional CDFs for the [Tamiami] metric have been presented in Figure 5-10, which show that the degree of skewness in the output CDF is related to the behavior of outliers at the time slice where the CDF is extracted. Thus, it can be expected that a FOSM analysis corresponding to the [Alt1] time slice would produce better agreement with MCS results than that for the [Alt2] time slice. In other words, the degree of agreement between the MCS and FOSM analyses depends on the degree of skewness of the output, or equivalently, the behavior of outliers beyond the $5^{th}/95^{th}$ percentile bounds in the horsetail plot.

6.0 UNCERTAINTY IMPORTANCE ANALYSIS

6.1 Introduction

Classical sensitivity analysis involves quantification of the change in a model output corresponding to a change in one or more of the model inputs. In the context of probabilistic models, however, sensitivity analysis takes on a more specific definition, namely, ranking and quantifying the contribution from individual input parameters to the uncertainty (the spread or variance) of model predictions (Helton, 1993). This is sometimes referred to as global sensitivity analysis or uncertainty importance analysis to distinguish it from the classical (local) sensitivity analysis measures typically obtained as partial derivatives of the output with respect to inputs of interest (Saltelli et al., 2000).

The contribution to output uncertainty (i.e., variance) by an input is a function of both the uncertainty of the input variable and the sensitivity of the output to that particular input. In general, input variables identified as important in global sensitivity analysis have both characteristics; they demonstrate significant variance and are characterized by large sensitivity coefficients. Conversely, variables that do not show up as important per these metrics are either restricted to a small range in the probabilistic analysis, and/or are variables to which the model outcome does not have a high sensitivity.

For this study, the goal of sensitivity analysis is to answer questions, such as:

- Which uncertain variables have the greatest impact on the overall uncertainty in probabilistic model outcomes?
- What are the key drivers of extreme values in the output metrics?

The analysis of the NSRSM results uses regression-based analyses and classification tree analyses to answer these questions, respectively. More explanation of these methods is presented in the following sections. The analyses are carried out using results from the probabilistic NSRSM calculations. The randomly sampled inputs considered in each of the realizations are treated as independent variables and the metrics computed from these inputs are treated as dependent variables.

6.2 Stepwise Rank Regression Analysis

6.2.1 Background

The computational scheme for the regression-based sensitivity analysis consists of two steps (Helton, 1993): (1) fitting a linear response surface between the output and the input variables and (2) performing sensitivity analysis on this "surrogate" model. Note that a multidimensional linear approximation for the model is a pre-requisite for this analysis. For models with nonlinear input-output dependencies, rank transformation has been reported to be a simple and effective linearizing technique when the output is a monotonic function of the inputs (Iman and Conover, 1979). Except as indicated, the theoretical discussion that follows applies equally to the cases of "raw" data and "rank-transformed" data.

Stepwise Regression Basics

The starting point for regression-based global sensitivity analysis is a multivariate linear rank regression model of the form:

$$\hat{y} = b_0 + \sum_j b_j x_j \tag{6-1}$$

where $\hat{y}$ denotes the fitted rank-transformed output, the x_j are the rank-transformed input variables of interest and the b_j are the unknown coefficients (Helton, 1993). The regression coefficients are generally fitted using a forward stepwise regression procedure (Draper and Smith, 1981) rather than treating all the independent variables in a single model. In this stepwise approach, a sequence of regression models is constructed starting with the input variable that explains the largest amount of variance in the output, i.e., the variable that has the highest simple (Pearson product-moment) correlation coefficient, *SCC*, with the output. The *SCC* is given by (Helton, 1993):

$$SCC[y, x_k] = \frac{\sum_j (x_{k,j} - \bar{x}_k)(y_j - \bar{y})}{\left[\sum_j (x_{k,j} - \bar{x}_k)^2 \sum_j (y_j - \bar{y})^2\right]^{1/2}} \tag{6-2}$$

The *SCC* provides a measure of the degree to which the input variable of interest and the predicted output change together, and quantifies the strength of linear and monotonic association between the input-output pair.

At each successive step in the regression modeling process, the variable that explains the largest fraction of unexplained variance from the previous step is included. This is the variable with the largest absolute value of the partial correlation coefficient, *PCC*, which measures the correlation between the output and the selected input variable after the linear influence of the other variables has been eliminated (Draper and Smith, 1981).

The model generated at every step is tested to ensure that each of the regression coefficients is significantly different from zero. A partial **F**-test is used to reject the hypothesis that a regression coefficient is zero, at a 100(1 - α) percent confidence level, where α is prescribed by the analyst (Draper and Smith, 1981). The test is implemented in two stages. First, a variable selected for entry via the *PCC* criterion is tested for its significance before it is admitted into the model. Second, after the model is built at that step, each of the variables in the model is tested for significance. If some variables are found to be insignificant, then the "most insignificant" variable is dropped and the model is built again.

The sequential dropping of the variables judged to be not significant and rebuilding the model continues until all the variables in the model become significant at the prescribed levels (α). The significance levels (α) are prescribed separately for the entering and departing variables to avoid possible looping where the same variable can enter and depart from the model. The significance level (α) for the departing variables is generally set larger than the entering variable's. Note that the need for dropping a variable generally arises only in the cases when the input variables are strongly correlated. This also ensures that multicolinearity is not present in the final model. While multicolinearity might not a significant issue for the model predictions, it can have detrimental effects on the value of the fitted coefficients.

The stepwise regression process continues until the input-output model contains all of the input variables that explain statistically significant amounts of variance in the output (i.e., no more variables are found with a statistically significant regression coefficient). Alternatively,

model overfitting can be diagnosed using a statistical metric such as PRESS (predicted error sum of squares) that assesses the tradeoff between the improvement in goodness-of-fit versus increase in the number of free parameters (Saltelli et al., 2000).

Uncertainty Importance Measures

Once the regression model building is complete, several metrics can be used as measures of uncertainty importance. The first of these is simply the order of entry into the regression model. Another commonly used measure is the standardized regression coefficient, *SRC*, defined for variable *j* as:

$$SRC_j = \frac{b_j \sigma(x_j)}{\sigma(y)} \tag{6-3}$$

where $\sigma(x_j)$ denotes the sample standard deviation of the uncertain input x_j and $\sigma(y)$ denotes the sample standard deviation of the output y (Helton et al., 1991). Obviously, the original x_j and y values are used here and not their rank-transformed values. The *SRC*s can be considered as regression coefficients that would be obtained from a regression analysis with the input and output variables normalized to zero mean and unit standard deviation. For uncorrelated inputs, the importance ranking with *SRC*s is identical to that with *SCC*s.

A related measure of importance is the fractional contribution to variance, *FCV*. When a linear additive input-output model is built with uncorrelated inputs, the goodness-of-fit of the model can be expressed as (Helton et al., 1991):

$$R^2 = \sum_j SCC^2[y, x_j] \tag{6-4}$$

where R^2, the coefficient of determination, denotes the fractional variance in y explained by the model. Thus, the term $SCC^2[y,x_j]$ can be interpreted as the fractional variance in y explained by the j-th independent variable. As can be easily ascertained, both *SCC* and *FCV* yield the same order of importance for the uncertain inputs that is also consistent with the *SRC* based importance ranking for uncorrelated inputs.

6.2.2 Stepwise Regression Results

Stepwise regression analyses were carried out using results from the probabilistic NSRSM simulations. mCalc was used to perform the majority of the stepwise regression calculations. The analysis was carried out with rank transformed data, using each metric as the dependent variable and the uncertain inputs as independent variables. The regression process was terminated when the regression coefficients were statistically indistinguishable from zero at the 95 percent significance level.

Table 6-1 shows the results of the stepwise regression for metric [25492stage]. A total of five variables were admitted to the regression model, corresponding to a final R^2 equal to 0.819. This means that roughly 82 percent of the variance in [25492stage] can be explained by the regression model. The most important variable, [KVEG511], accounts for 46 percent of the variance explained by the input-output model. Note that this is 46 percent of the *explained* variance, i.e. 0.379 / 0.819 * 100. This input parameter represents the vegetation coefficient for the Ridge and Slough Marsh, so we can conclude that the uncertainty in the stage at location 25492 is strongly influenced by factors controlling evapotranspiration for this vegetation type.

The second variable admitted to the regression model is [ALPHA511], which accounts for an additional 20 percent of the explained variance. This parameter is the Manning's coefficient for the Ridge and Slough Marsh, indicating that factors controlling surface water flow (and depth) also have some effect on the uncertainty in the stage at location 25492. The [TOPOSELECT] variable accounts for an additional 13 percent of the variance in the output metric. This indicates that the local topography (see Figures 2-5 and 2-6 for the area where topography varied among the three cases) also has a significant effect on the uncertainty in the stage at location 25492. The other variables in the regression model are only of marginal importance, as indicated by the relatively small increases in the R^2 values in Table 6-1. Note that the smaller increases in R^2 values correspond to small absolute values of SRC.

Table 6-1 Stepwise-Regression Analysis Results for metric [25492stage].

Rank	Variable	R^2	SRC
1	KVEG511	0.379	-0.632
2	ALPHA511	0.545	0.425
3	TOPOSELECT	0.653	0.322
4	DETENT511	0.750	0.315
5	KVEG712	0.819	-0.264

Another tool for investigating the relationship between model output and key uncertain inputs is the use of scatter plots. Figure 6-1 shows the scatter plots between [25492stage] and the first four variables added to the regression model. The red line shown in each scatterplot is a LOESS (local regression smoothing) trendline. These trendlines are useful when cluster densities or other factors make interpretation of trends difficult. Note that the sign of the SRC should indicate the direction of the trend in the scatterplot, i.e. a negative SRC indicates a decreasing trend in the scatterplot, as shown with [KVEG511].

Figure 6-1 shows the influence of [KVEG511], with a strong, negative, and nearly linear input-output relationship. The input-output relationship between [25492stage] and [ALPHA511] is less pronounced, with an upward trend showing extensive scatter. The other two variables show weaker positive trends. The plot [TOPOSELECT] demonstrates the difficulty in visually interpreting results from categorical variables. In this case, the smoothing trendline is especially helpful in showing the direction of correlation.

Table 6-2 shows the results of the stepwise regression for metric [25087]. A total of 3 variables were added to the regression model, resulting in a final R^2 of 0.982. The top-ranked variable, [KVEG712], dominates the regression, accounting for 97% of the explained variance. The other two variables are of marginal additional importance. As expected, because location 25087 is in the Mesic Pine Flatwood land cover type, parameters from this land cover type dominate the regression model.

The scatterplots for [25087stage] are shown in Figure 6-2. Note the very strong negative trend shown in the plot for [KVEG712], with nearly horizontal (indicating little correlation) trends for the other two variables.

Table 6-2 Stepwise-Regression Analysis Results for metric [25087stage].

Rank	Variable	R^2	SRC
1	KVEG712	0.971	-0.981
2	ALPHA712	0.981	0.099
3	DETENT712	0.982	0.044

Table 6-3 shows the results of the stepwise regression for metric [Tamiami]. A total of 5 variables were added to the regression model, resulting in a final R^2 of 0.871. As would be expected, because the Tamiami transect is in the Ridge and Slough Marsh land cover type, Ridge and Slough Marsh parameters comprise the first three variables added to the regression model. The marginal contribution of the remaining variables is evident in their low SRC values, of approximately 0.2 or less. As with the stage metric in this land cover type, the vegetation coefficient is the most important variable, followed by Manning's *n*.

Figure 6-3 shows the scatterplots corresponding to the variables in the regression. The first scatterplot shows tightly clustered values around the trendline, along with a few higher values of [Tamiami] that do not follow the trend as closely. These higher values are among those that comprise the flat section in the upper decile for the CDF of this metric (Figure 5-8). The scatterplot for the second variable shows a nonlinear, but monotonically decreasing trend. The rest of the scatterplots show weaker correlation.

Table 6-3 Stepwise-Regression Analysis Results for metric [Tamiami].

Rank	Variable	R^2	SRC
1	KVEG511	0.557	-0.679
2	ALPHA511	0.742	-0.428
3	KVEG712	0.810	-0.259
4	DETENT511	0.861	0.214
5	XD511	0.871	0.103

Table 6-4 shows the results of the stepwise regression for metric [T712_East]. A total of 4 variables were added to the regression model, with a final R^2 of 0.943. [KVEG712] comprises about 66% of the explained variance, followed by [ALPHA712] with an additional 28% of the explained variance. The rest of the variables are of marginal importance. Figure 6-4 is consistent with this result, where the [KVEG712] scatterplot shows another strong, monotonic

trend, and the [ALPHA712] plot also shows a significant trend. The other variables show weaker correlation in their respective plots.

Table 6-4 Stepwise-Regression Analysis Results for metric [T712_East].

Rank	Variable	R^2	SRC
1	KVEG712	0.620	0.800
2	ALPHA712	0.889	0.508
3	DETENT712	0.941	-0.228
4	ALPHA511	0.943	0.048

These regression results indicate that the vegetation coefficient and Manning's *n* are consistently the most important variables for explaining the uncertainty in all of the output metrics, with some contribution from topography for the [25492stage] metric.

6.3 Classification Tree Analysis

6.3.1 Background

Although linear regression is routinely used for analyzing the entire spectra of output data, specialized approaches may be required for examining small subsets (e.g., top and bottom deciles) of the output. To this end, classification tree analysis can provide useful insights into what variable or variables are most important in determining whether outputs fall in one or the other (extreme) category (Breiman et al., 1984). Traditional applications of classification trees have primarily been in medical decision making and data mining for social sciences. Mishra et al. (2003) describe an application of the methodology to a Monte Carlo simulation-based model for predicting performance of a potential nuclear waste repository. Compared to linear regression modeling, tree-based models are attractive because they can handle more general interactions between predictor variables. In other words, the representation of interaction between input variables need not be restricted to simple forms such as x1*x2 or x1/x2 as is typically done in standard multivariate linear (or linearized) regression modeling. Also, results of the tree-based modeling are insensitive to monotonic transformations (e.g., logarithmic or rank) of the input variables, which makes the analysis more robust.

A binary decision tree is at the heart of classification tree analysis. The decision tree is generated by recursively finding the variable splits that best separate the output into groups where a single category dominates. The degree by which a single category dominates is called the split "purity". For each successive fork of the binary decision tree, the algorithm searches through the variables one by one to find the purest split within each variable. The splits are then compared among all the variables to find the best split for that fork. The process is repeated until all groups contain a single category, or a specified level of purity is reached for all groups. In general, the variables that are chosen by the algorithm for the first several splits are most important, while less important variables are normally involved in the splitting near the terminal nodes of the tree.

The tree-building methodology used in the current study is based on a probability model approach. Classifiers at each node are selected based on an overall maximum reduction in impurity, for all possible binary splits over all the input variables. The impurity at a given node A is based on the Gini index (Breiman et al., 1984), which for the two class case reduces to:

$$I_A = 2 p_{1A} p_{2A} \tag{6-5}$$

where p_{1A} and p_{2A} are the estimated probabilities of classes 1 and 2, respectively, at node A. We do not know the probabilities, but can estimate them from the proportion at a node, i.e.:

$$p_{1A} = \frac{n_{1A}}{n_A} \tag{6-6}$$

The decrease in impurity for a given split of node A into nodes L and R (left and right) is

$$\Delta I = I_A - p_L I_L - p_R I_R \tag{6-7}$$

where p_L and p_R are the proportions of the cases that go to L and R, respectively.

The classification tree is built by successively taking the maximum reduction in purity over all the allowed splits of the leaves to determine the next split. Termination occurs when the number of cases at a node drops below a set minimum, or when the maximum possible reduction in purity for splitting a particular node drops below a set minimum.

For the current study, classification tree analysis was completed for the 200 realization case for each output metric, based on all input metrics. The output metric was divided into "high" and "low" categories, where the "high" category consisted of the 50 results in the upper quartile (75 – 100 percentile) and the "low" category consisted of the 50 results in the lower quartile (0 – 25 percentile). The statistics software, R (Venables et al., 2006), was used to create the classification trees based on these categories.

6.3.2 Classification Tree Results

Figure 6-5 shows the classification tree for metric [25492stage]. The first split occurs on the most important parameter, [KVEG511]. Splitting the data at [KVEG511] = 0.8409 resulted in 32 values in the "low" category on the right and a mixture of "high" and "low" on the left. Further splits on the [ALPHA511] variable creates nearly pure final nodes, with 49 "high" on the left and 18 "low" on the right (along with one "high"). So this analysis indicates that [KVEG511] is the most important input parameter for determining extreme results in the metric [25492stage], followed by [ALPHA511].

Another way of showing this result is via a partition plot, as shown in Figure 6-6. A partition plot is simply a scatterplot of the top two variables of the tree, with different symbols for the two categories, "high" and "low". One horizontal and one vertical line correspond to the location of the splits for the input variables. The main utility of a partition plot is to display the clustering of outcomes (if any) in the parameter space. This helps provide a visual interpretation of the decision rules generated by the classification tree algorithm. Figure 6-6 shows the clustering of the "high" category in the upper left quadrant based on the two splits.

As noted in the regression section, for many of the metrics a single variable is dominant in the explained variance. As a result, we would expect the classification trees to be dominated by the first split, with further splits either having negligible impact or being altogether unnecessary. The variable involved in the first split will correspond to the first variable added in the regression model discussed in the previous section. Figure 6-7 shows the classification tree for metric [25087stage]. Note that the first split on [KVEG712] produces pure nodes, as might be expected given the strength of the association between the variable and the metric shown in

Figure 6-2. Because there is only one split, the partition plot does not provide any useful visualization and is therefore not presented here.

Figure 6-8 shows the classification tree for metric [Tamiami], where a single split on parameter [KVEG511] creates most of the separation between the categories. The partition plot in Figure 6-9 shows this effect, with most of the division between the categories occurring across the vertical line at [KVEG511] = 0.8169.

Figure 6-10 shows the classification tree for [T712_East]. As with the [Tamiami] result, most of the separation occurs on the first split with variable [KVEG712]. The partition plot in Figure 6-11 illustrates this, with most of the separation in the two classes occurring across the vertical line at [KVEG712] = 0.6329.

6.4 Discussion

In the current work, we have applied three techniques for evaluating the sensitivity of output metrics to input parameters: singular value decomposition, stepwise rank regression, and classification tree analysis. Singular value decomposition is a "local" sensitivity analysis, providing the relative influence of parameters at a single point in the parameter space. The strengths of the SVD-based approach lie in potential computation efficiency and providing insight into parameter interactions in their influence on the output metrics. Regressions analysis is a "global" sensitivity analysis, in that the full parameter space defined by the input distributions is analyzed. It is dependent on results from an MCS, and should provide a robust result. Also, categorical variables (such as [TOPOSELECT]) that are non-differentiable can often be included in the sensitivity analysis by using this MCS/stepwise regression combination. Classification tree analysis also requires the results from an MCS, but explores the extremes of the output metrics. This technique can also be used with discrete parameters and provides simple parameter rules for determining what governs extreme results in the output metrics. When two or more variables have comparable explanatory power, partition plots are often useful for visualizing what combinations of input values would produce different categorical outcomes (e.g., fit v/s misfit).

In some cases, these three techniques for sensitivity analysis can provide different insights for analysis of a particular model. For the current case, the results from the three analyses are very consistent. . For all techniques and all metrics, the vegetation coefficient and Manning's *n* were determined to be the primarily influential parameters. The reasons for such consistent behavior can be explained as follows. The SVD-based sensitivity analyses essentially capture the relative importance of the locally linear input-output sensitivity coefficients at the reference point. The stepwise regression based sensitivity analyses capture the relative importance of rank-transformed input-output dependencies over the sampled range. Thus, as long as there are no important non-monotonic input-output relationships, and the rank transformation is successful in linearizing input-output behavior, the importance ranking from the SVD-based analysis and stepwise-regression based analysis should be consistent.

On one hand, it is encouraging to have consistent results from different techniques, because it makes a convincing case that the results are robust. On the other hand, it does not go as far in demonstrating the unique insights that can be provided by each of the techniques. We should consider whether the effort to apply all three techniques is justified in all cases. As noted before, the dominance of a single variable in the explanation of variance for a particular metric creates a situation where classification tree results may add little to the stepwise regression results. In the future, it may be recommended that if the results from the regression analysis indicate single variable domination for a metric, especially when these results are consistent with the SVD results, the classification tree building exercise will be unnecessary.

7.0 SUMMARY AND CONCLUSIONS

This report describes the application of a sensitivity and uncertainty analysis methodology for the NSRSM using the uncertainty analysis software mCalc. The goal is to implement a systematic methodology that includes uncertainty characterization, uncertainty propagation and uncertainty importance (global sensitivity) analysis.

The study begins with an identification of a limited number of key uncertain inputs and output metrics of interest. Five uncertain inputs (i.e., Manning's *n*, detention storage, vegetation crop coefficient, extinction depth and storage coefficient) are selected for two land cover conditions, (i.e., Ridge and Slough Marsh and Mesic Pine Flatwood). Uncertainty in these parameters is characterized using a limited amount of literature data, model calibration values from analogous regions, and expert judgment. Uncertainty in topography is represented using a reference topographical map, and two extreme scenarios around the reference case.

The output metrics selected are average water level for a one week time window at two model cell locations and the average flow for a one week time window at two transect locations. The time windows were selected within the representative water year May 1992 – April 1993.

An SVD-based sensitivity analysis is carried out to examine the sensitivity between the selected output metrics, individually and in aggregate, and the selected uncertain inputs. The results provide insight into dominant parameter groups controlling model behavior in the vicinity of the reference point.

The mCalc uncertainty analysis software is employed next for uncertainty propagation using Monte Carlo simulation. Stability of model results is investigated using sample size of 100, 200, and 300 – with the 200 realization case being selected for detailed analyses. The probability distributions of the selected output metrics reveal a broad spectrum of responses – ranging from the symmetric to the skewed. In general, the coefficient of variation associated with model outcomes is less than 20%, with only a few exceptions. It should be noted that the observed skewness of the output is dependent to some extent on where the CDF is extracted from the horsetail plot.

The more computationally efficient FOSM approach is also employed to compute the mean and standard deviation of the output metrics. In general, estimates of the mean and standard deviation agree well between the MCS and FOSM approaches, with the most deviation for the [Tamiami] metric which displayed a particularly skewed CDF. An important issue is whether the selected output metrics can produce stable derivatives using a standard forward or central difference approximation, failing which the theoretical computational efficiency of the FOSM may not be realized.

Finally, uncertainty importance analysis of MCS results is carried out using stepwise regression analysis and classification tree analysis. Based on input-output regression modeling, the key variables dominating the uncertainty in output metrics are found to be vegetation coefficient, Manning's *n*, and to a lesser extent, topographic uncertainty. Note that this importance ranking could be influenced by the particular choice of a discrete representation of topographic uncertainty influencing a small spatial domain. In many cases, only one variable contributes significantly to the overall variance of the output.

In classification tree analysis, the input-output relationship is reanalyzed by examining the factors responsible for separating the extreme outcomes (top 25% and bottom 25%). The results are consistent with regression analysis, where the uncertainty in most output metrics are dominated by a single variable. Both of these techniques (stepwise regression and classification tree analysis) produce importance rankings that reflect a combination of input-output sensitivity and input uncertainty. On the other hand, the importance ranking from SVD-factorization is only dependent on input-output sensitivity.

Recommendations based on this study can be summarized as follows:

- Additional effort is required for a more comprehensive characterization of parameter uncertainty. The observed dominance of a single variable vis-à-vis several output metrics could be an artifact of the subjectivity in assigning distributions. An example is the discrete representation for topographic uncertainty, which could be better characterized with an ensemble of equi-probable 2-D geostatistical fields.

- The selection of output metrics, and the time slices where the CDFs are extracted, should take into consideration the intended use of the probabilistic model and the associated decision variables. In a time-dependent model, the horsetail plot produces a general indication of uncertainty in computed outcomes. More detailed insights can be obtained from the analysis of CDFs at various time slices. The choice of such time slices should be consistent with the decision variables of interest (e.g., maximum stage/average flow for representative period).
- When using techniques that require calculation of the Jacobian, one should start with some exploratory runs that ensure the perturbation responses are monotonic and approximately linear. Otherwise, one should consider using a multipoint derivative algorithm. Another option is to carefully select an output metric so as to help improve the behavior of the response.
- When using multi-point derivatives, the application of the FOSM method for predicting output uncertainty should be reconsidered, since the computational efficiency of the method suffers with such a derivative calculation strategy.
- The results of the stepwise regression should be analyzed before completing a classification tree analysis. The fit of the regression model and the scatterplots should be examined to see if analysis of extremes would be useful. Otherwise, the classification tree analysis would basically duplicate the regression analysis.
- Overall, this work achieved its objective of demonstrating the utility of several sensitivity and uncertainty analysis techniques for application to the NSRSM model. We should be confident that future application of these techniques will aid any decision-making process that is contingent on results from the NSRSM model. Uncertainty analysis provides probabilistic results that allow quantitative risk-based decision-making by management. By identifying the few parameters that drive uncertainty in chosen metrics, sensitivity analysis allows resources to be more efficiently focused to improve reliability of future estimates. Plans for future work on probabilistic NSRSM modeling should consider recommendations made in this section, specifically regarding the comprehensive characterization of input

uncertainty, and selection of output metrics that are clearly related to management decision making.

8.0 REFERENCES

Ang, A.H.-S., and Tang, W.H. (1984). Probability Concepts in Engineering Planning and Design, vol. II, Decision, Risk and Reliability, John Wiley, New York.

Benjamin, J.R. and Cornell, C.A., 1970, Probability, Statistics and Decisions for Civil Engineers, Mc-Graw Hill, New York.

Breiman, L., J.H. Friedman, R.A. Olshen and C.J. Stone, 1984, Classification and Regression Trees. Wadsworth and Brooks/Cole, Monterey, CA.

Dettinger, M.D. and Wilson, J.L. (1981). "First Order Analysis of Uncertainty in Numerical Models of Groundwater Flow," Water Resources Research, **17**, No. 1, 149.

Doherty, J., 2004, PEST Model-Independent Parameter Estimation User Manual: 5th Edition, Watermark Numerical Computing, 336pp.

Draper, N.R. and H. Smith, 1981, Applied Regression Analysis, 2nd Ed., John Wiley, New York, NY.

Helton, J.C., 1993, "Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive waste disposal," Reliability Engineering & System Safety, 42 (2-3), 327-367.

Helton, J.C., J.W. Garner, R.D. McCurley and D.K. Rudeen, 1991, Sensitivity Analysis Techniques and Results for Performance Assessment at the Waste Isolation Pilot Plant, Sandia National Laboratories, Report SAND90-7103.

Iman, R.L., and W.J. Conover, 1979, "The use of rank transform in regression," Technometrics, 21 (4), 499-509.

James, A.L. and Oldenburg, C.M., 1997, "Linear and Monte Carlo Uncertainty Analysis for Subsurface Multiphase Contaminant Transport," Water Resources Research, **33**, No. 11, 2495.

Keefer, D. and S.E. Bodily, 1983. "Three-point approximations for continuous random variables," Management Science, 29, 595-609.

Lal, A.M.W., 1995, "Calibration of riverbed roughness," J. Hydraulic Engineering ASCE, 121 (9), 664-671.

McKay, M. D., W.J. Conover and R.J. Beckman, 1979, "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code," Technometrics, 21 (3), 239-245.

Mishra, S. and Parker, J.C., 1989, "Effects of Parameter Uncertainty on Predictions of Unsaturated Flow," Journal of Hydrology **108**, 19.

Mishra, S., N.E. Deeds and B.S. RamaRao, 2003, "Application of classification trees in the sensitivity analysis of probabilistic model results", Reliability Engineering & System Safety, 73, 123-129.

Morgan, M.G. and Henrion, M., 1990, Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis, Cambridge University Press, New York.

Saltelli, A., K. Chan and M. Scott, editors, 2000, Sensitivity Analysis, John Wiley, New York.

Trimble, P.J., 1995, An Evaluation of the Certainty of System Performance Measures Generated by the South Florida Water Management Model, MS Thesis, Florida Atlantic University, 245pp.

Tung, Y-K. and B-C. Yen, 2005, Hydrosystems Engineering Uncertainty Analysis, McGraw Hill, New York.

Venables, W. N., Smith, D. M., R Core Development Team, 2006, An Introduction to R, http://www.r-project.org, 99pp.

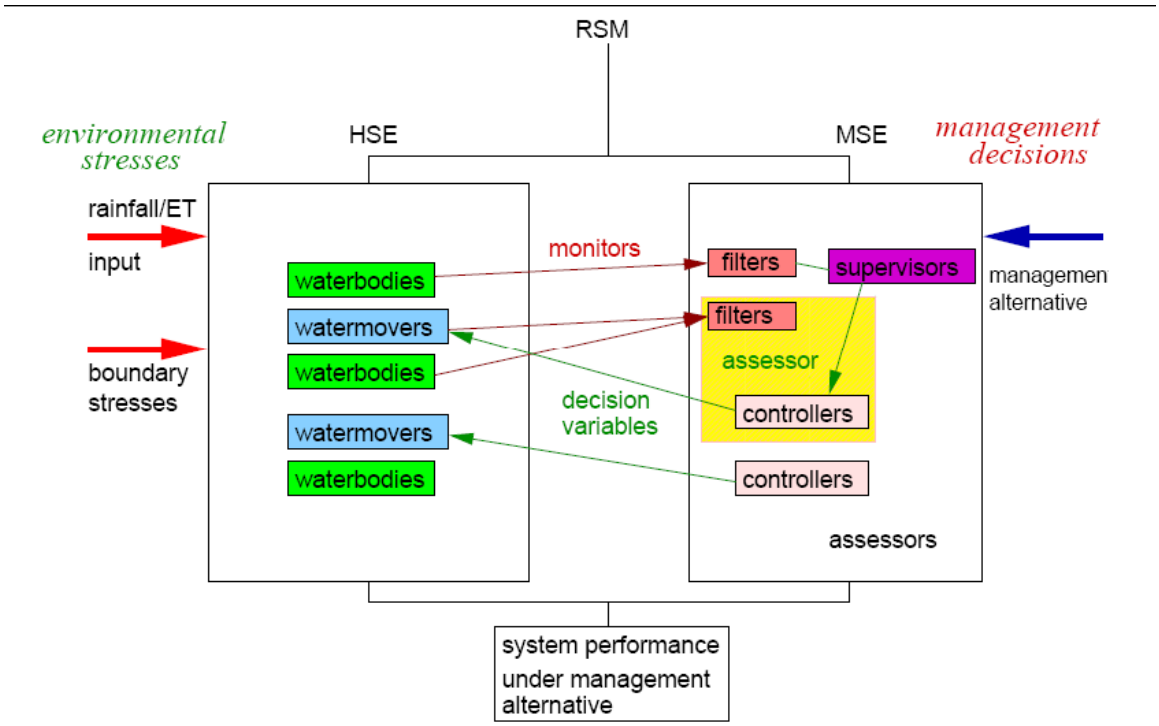


Figure 2-1 Structure of RSM.

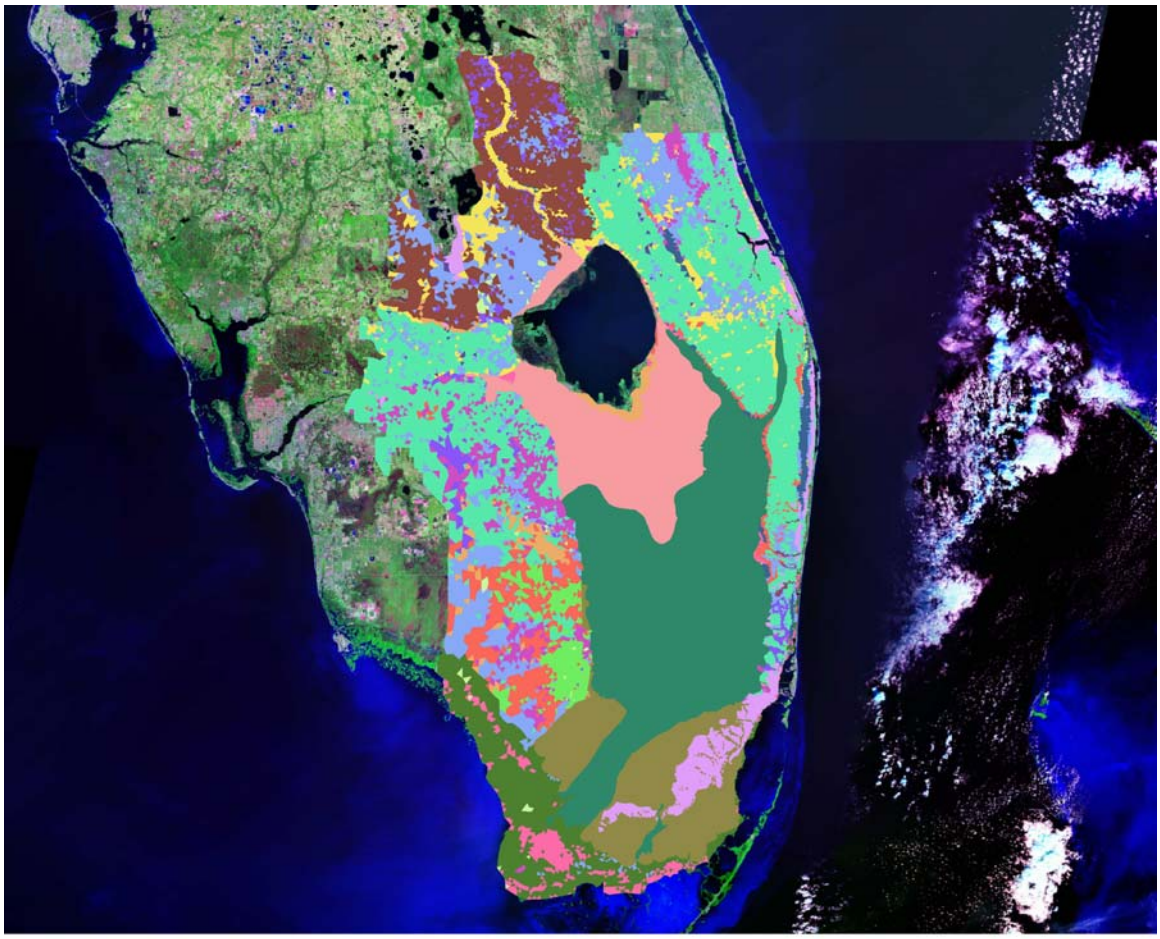


Figure 2-2 Predevelopment land use based on historical data.

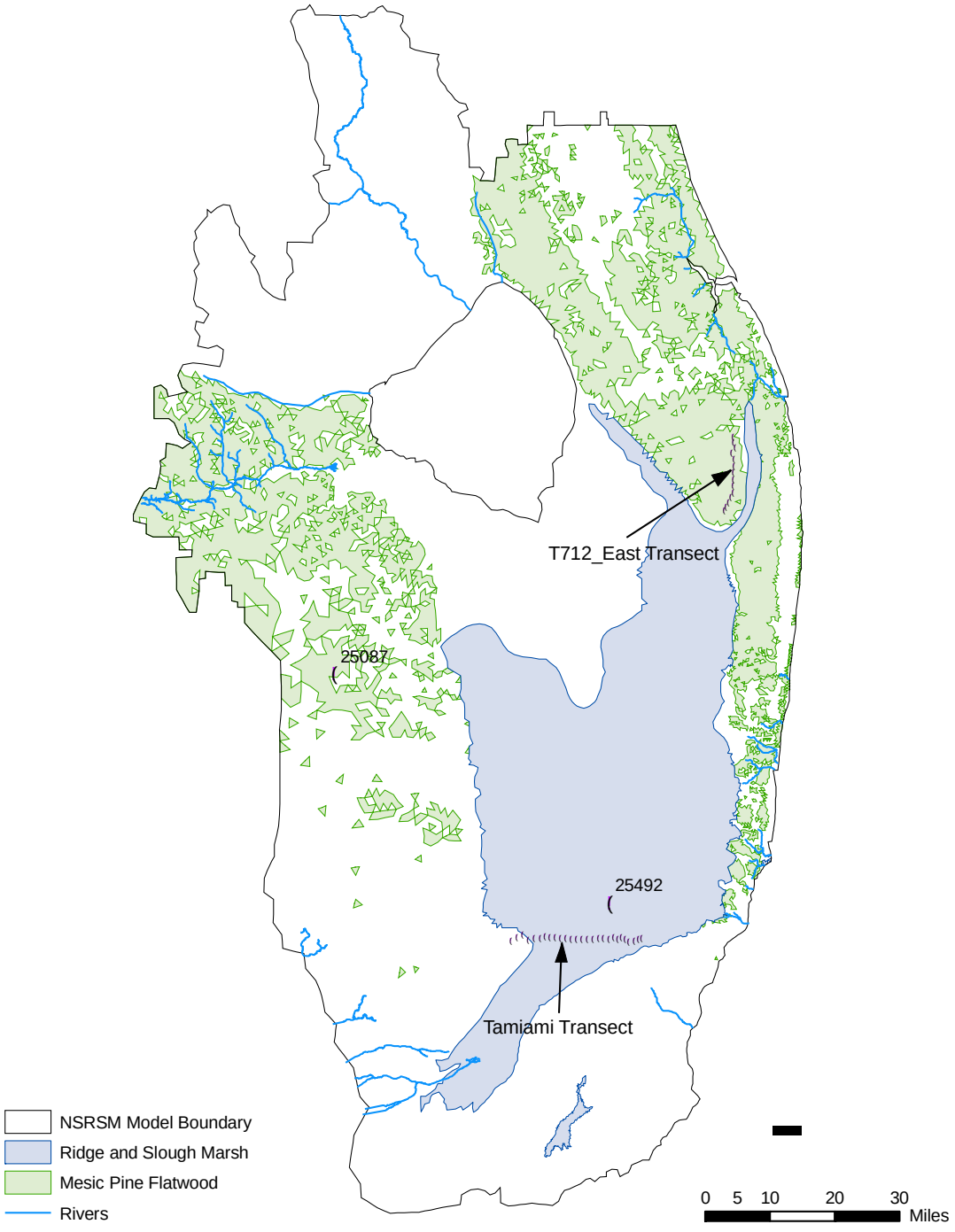


Figure 2-3 Land cover types chosen for parameter variation and locations of output metrics.

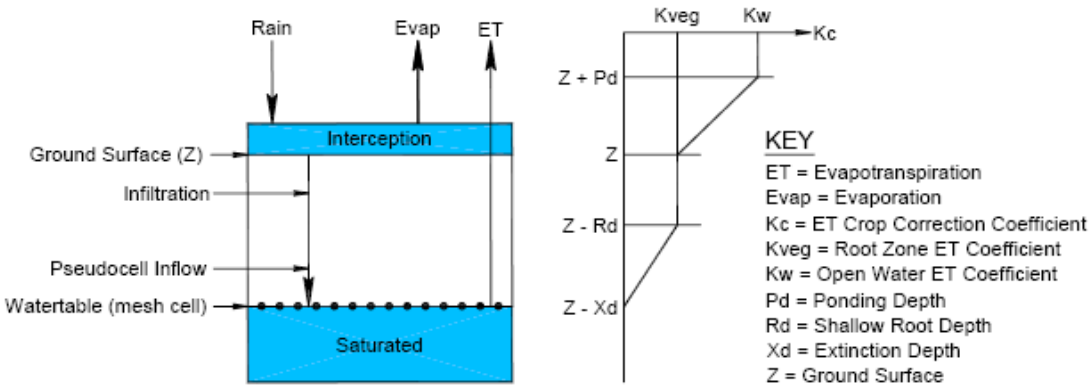


Figure 2-4 Modeling of ET in the RSM.

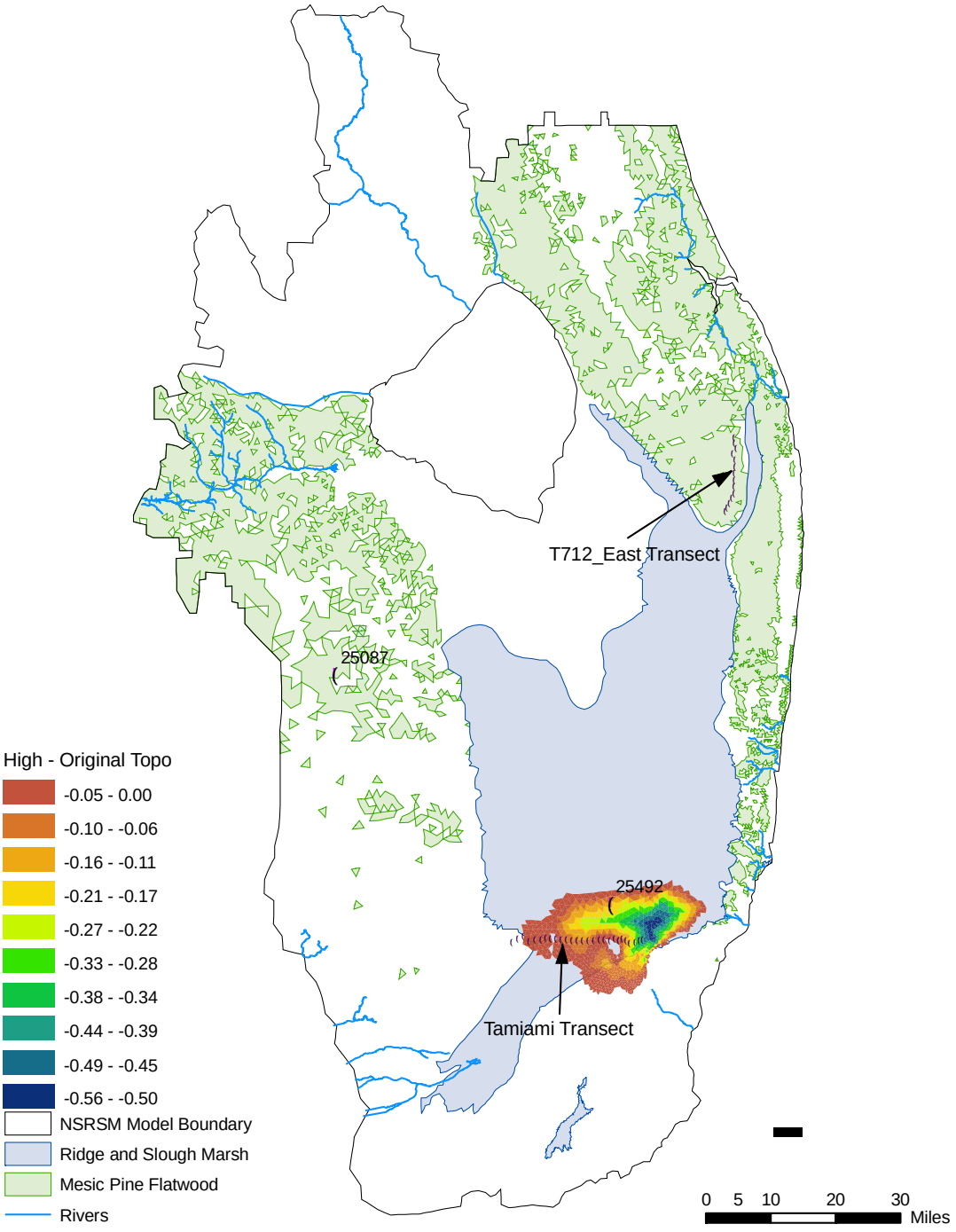


Figure 2-5 Difference between "high" topographic map and "base" topographic map.

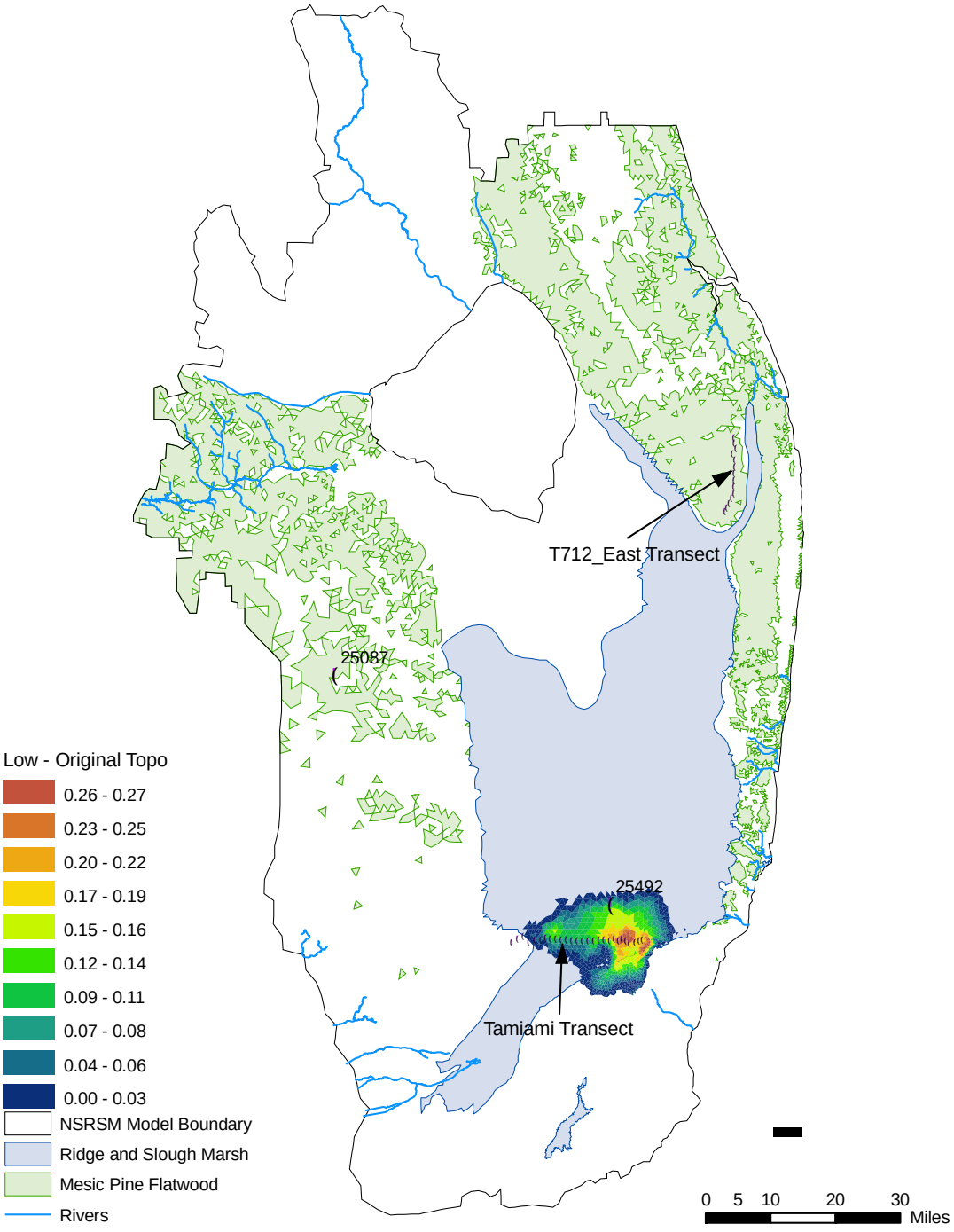


Figure 2-6 Difference between "low" topographic map and "base" topographic map.

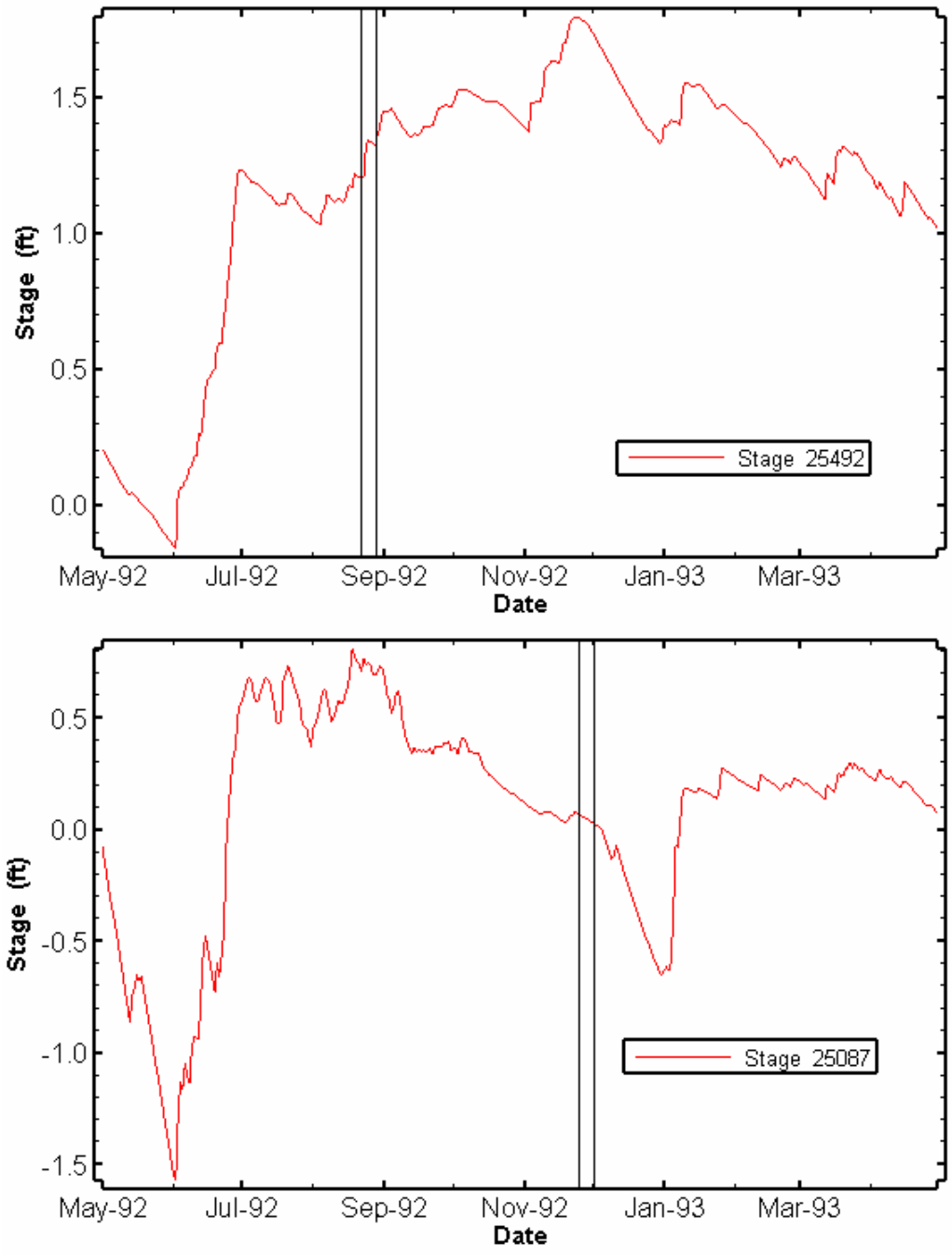


Figure 2-7 Water elevation hydrographs for locations 25087 and 25492 with averaging time window for sensitivity and uncertainty analysis.

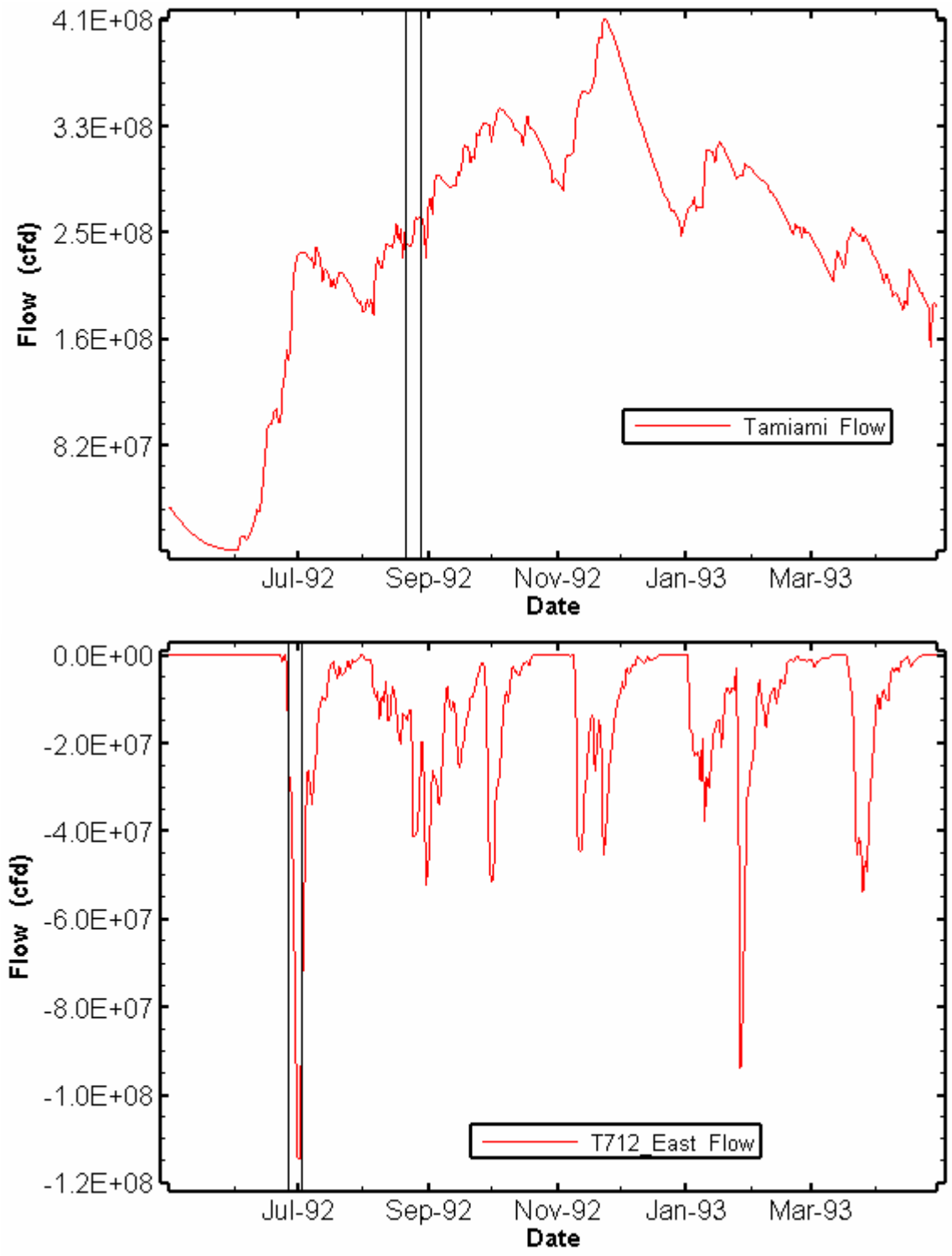


Figure 2-8 **Transect flow hydrographs for transects Tamiami and T712_East with averaging time window for sensitivity and uncertainty analysis.**

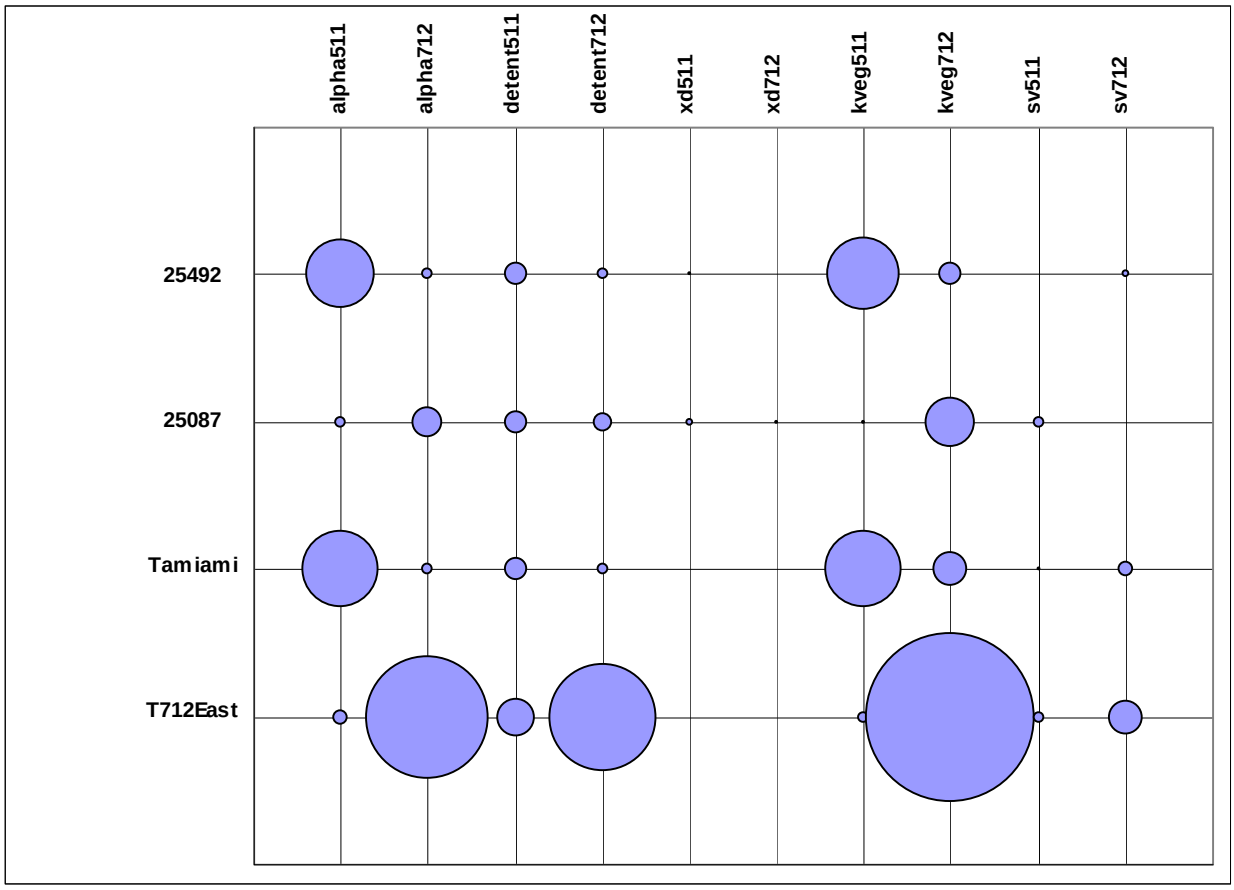


Figure 3-1 Bubble plot of the sensitivity matrix.

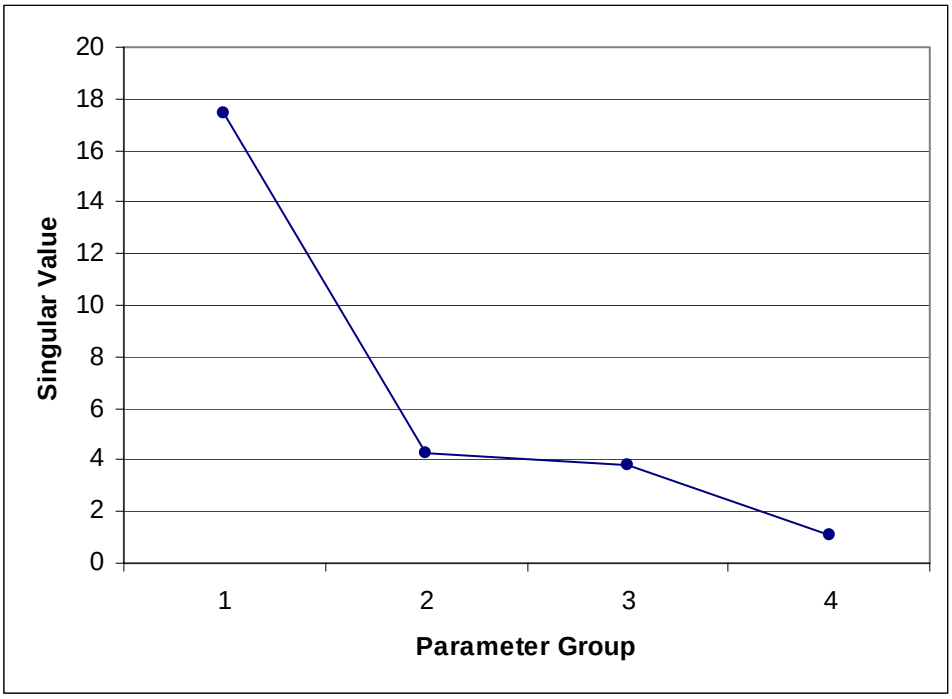


Figure 3-2 Singular values from the SVD decomposition.

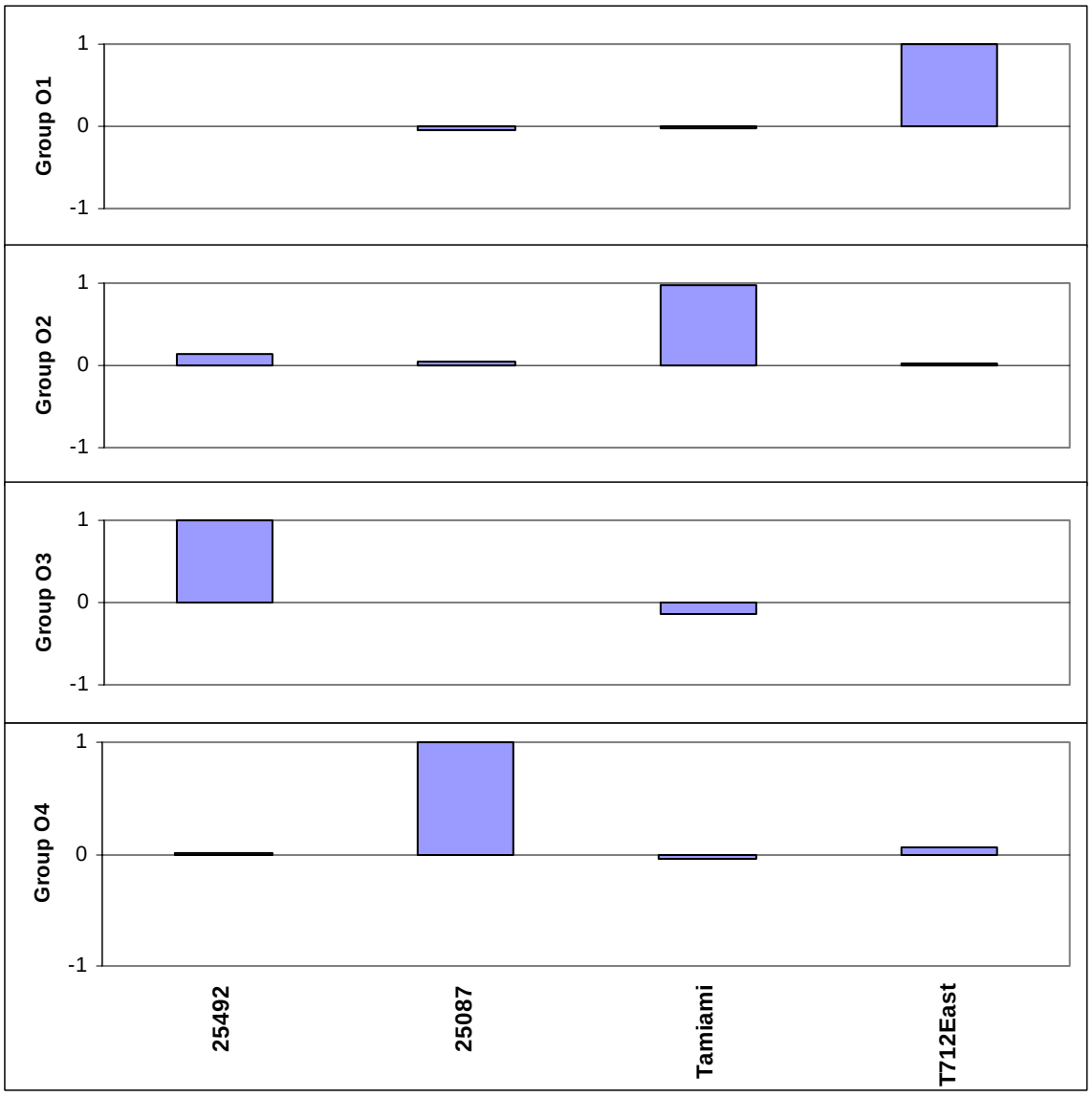


Figure 3-3 ***U*** **matrix elements showing linear coefficients of the output groups.**

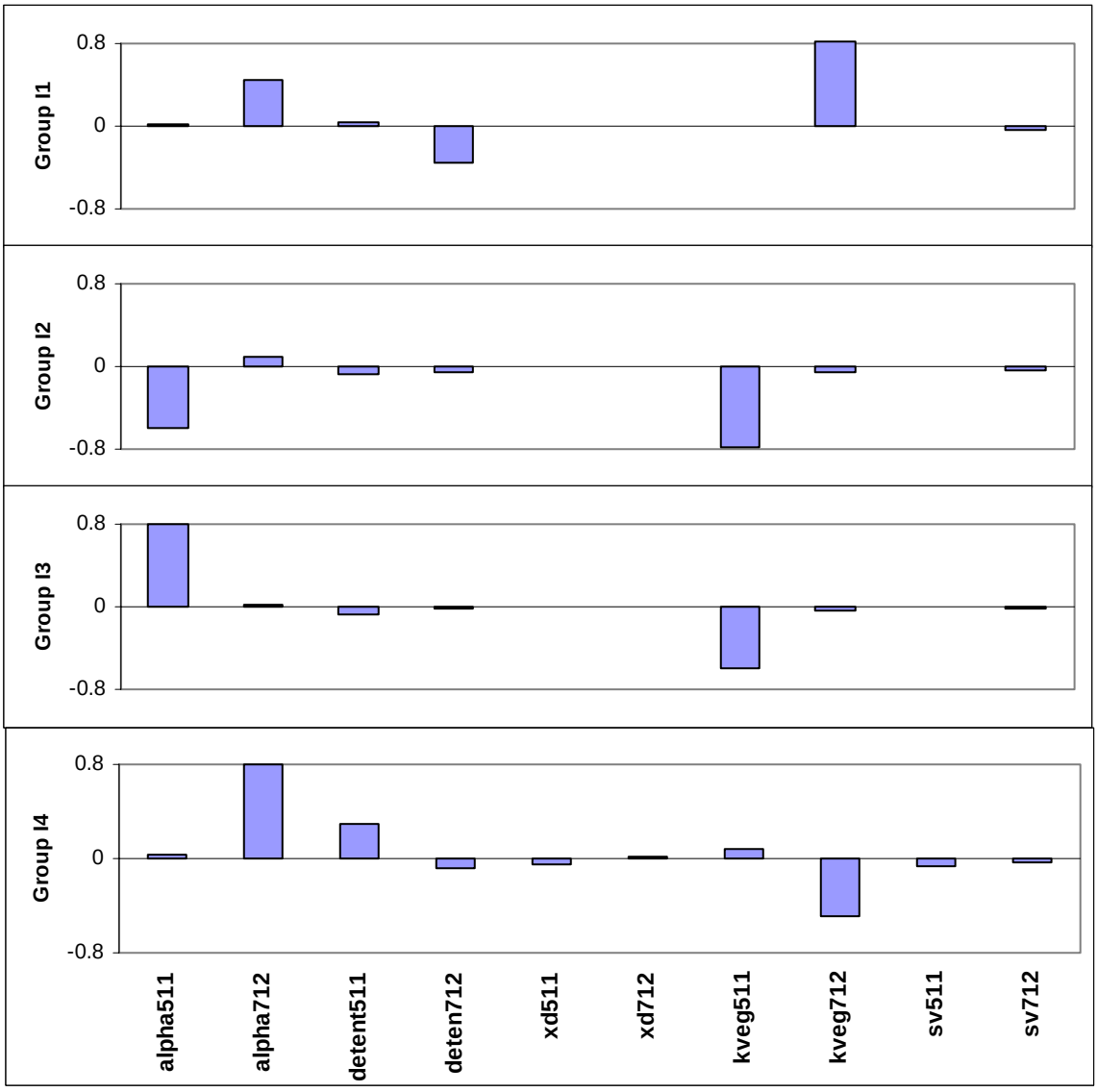


Figure 3-4 Elements of the V^T matrix showing linear coefficients of parameter groups.

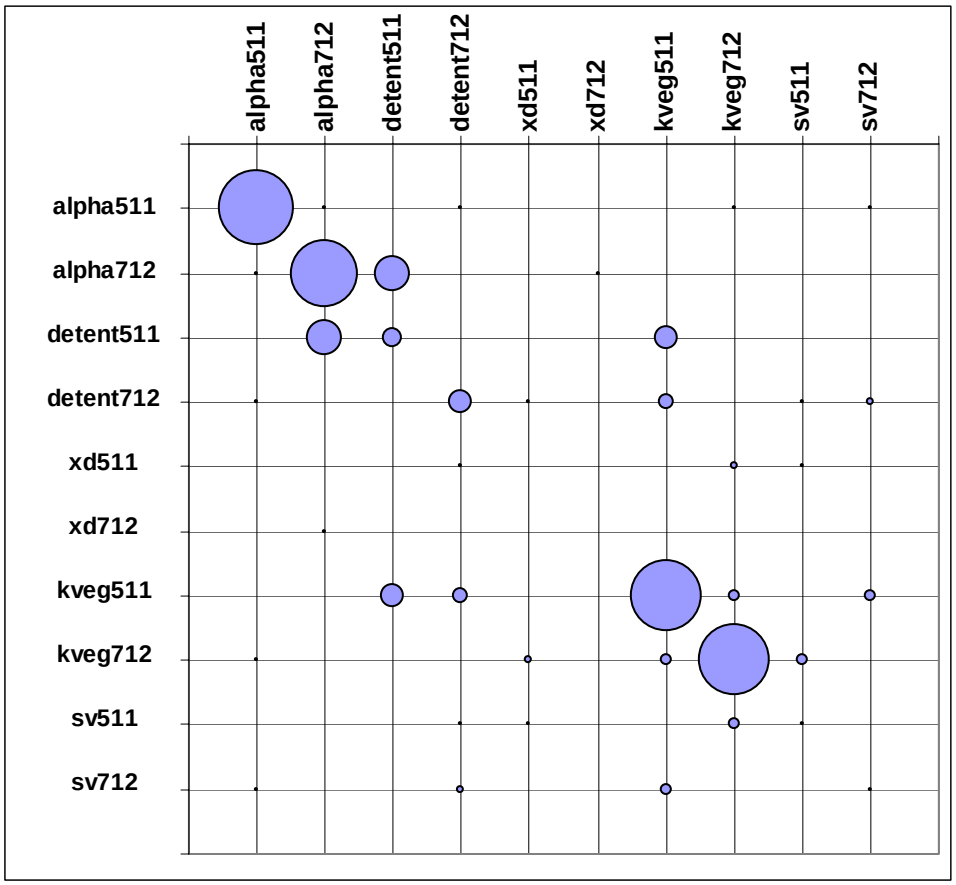


Figure 3-5 Resolution matrix from the SVD decomposition.

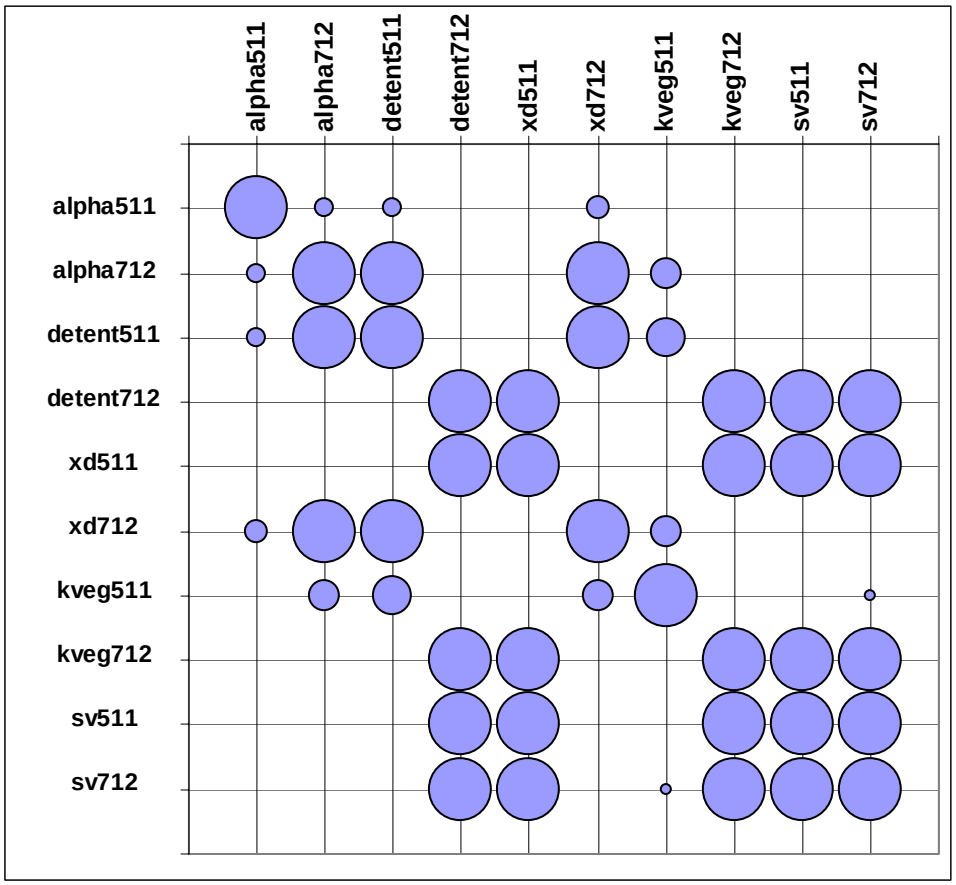


Figure 3-6 Correlation matrix from the SVD decomposition.

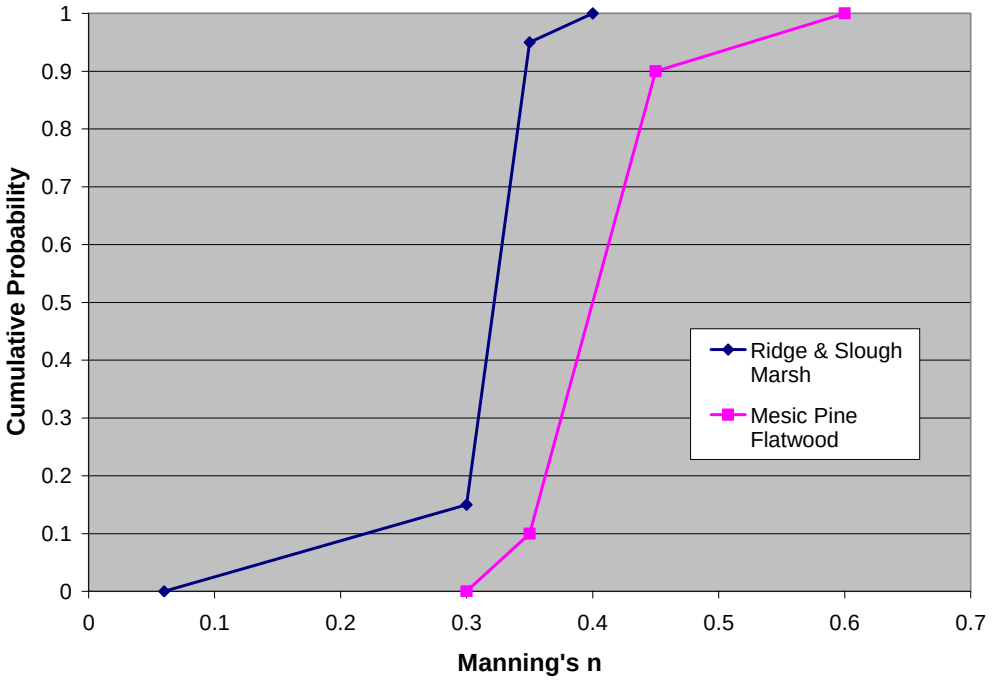


Figure 4-1 Cumulative distribution function (CDF) for Manning's *n*.

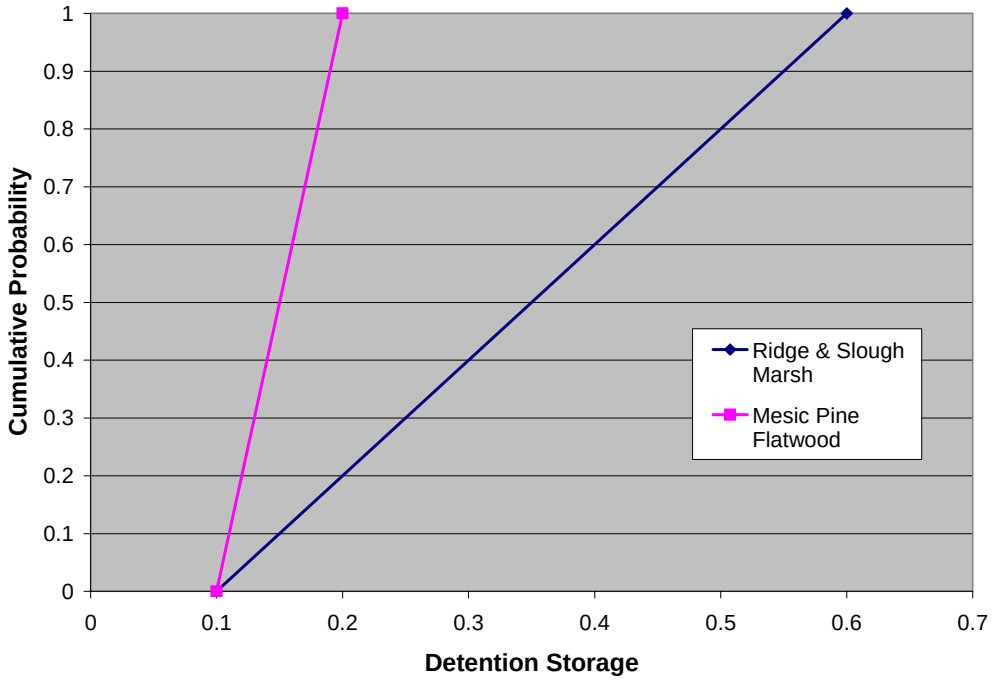


Figure 4-2 Cumulative distribution function (CDF) for detention storage.

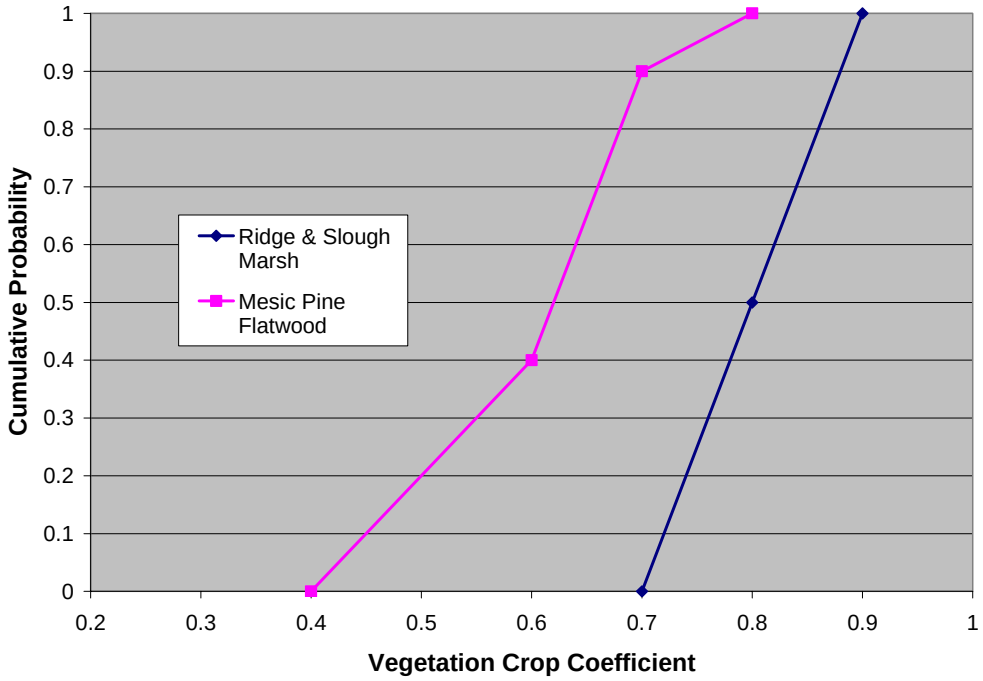


Figure 4-3 Cumulative distribution function (CDF) for vegetation crop coefficient.

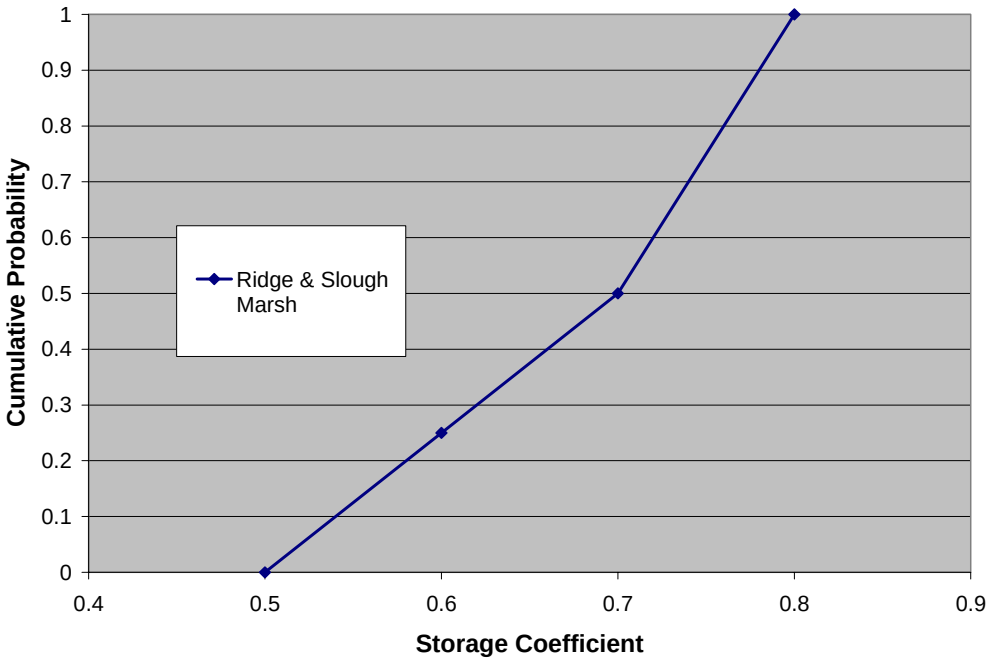


Figure 4-4 Cumulative distribution function (CDF) for storage coefficient.

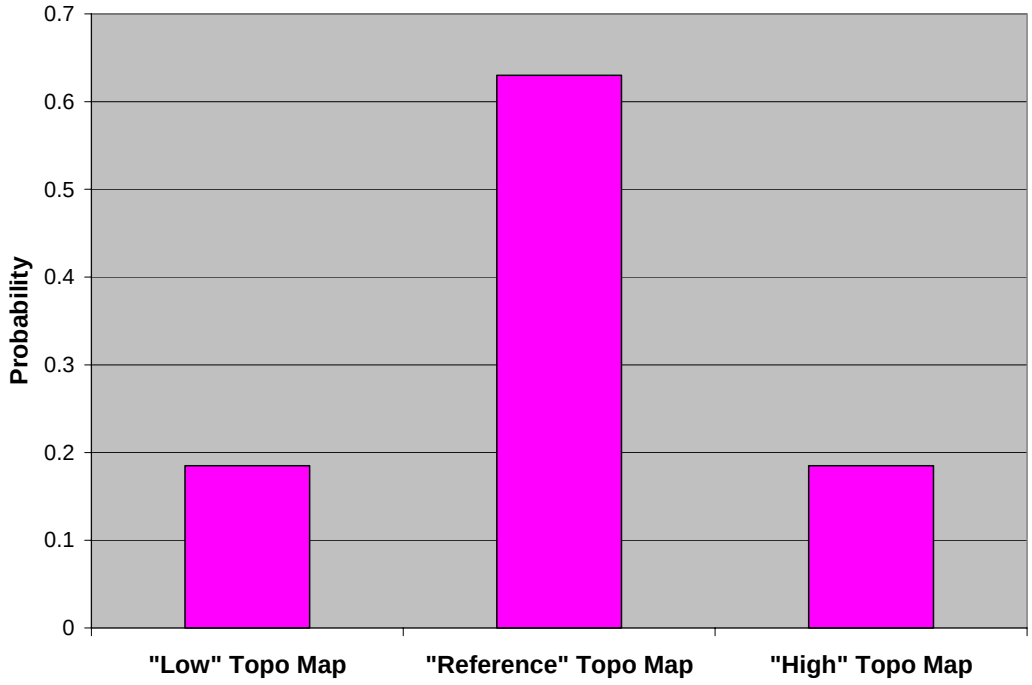


Figure 4-5 Probability density function (PDF) for topography indicator variable.

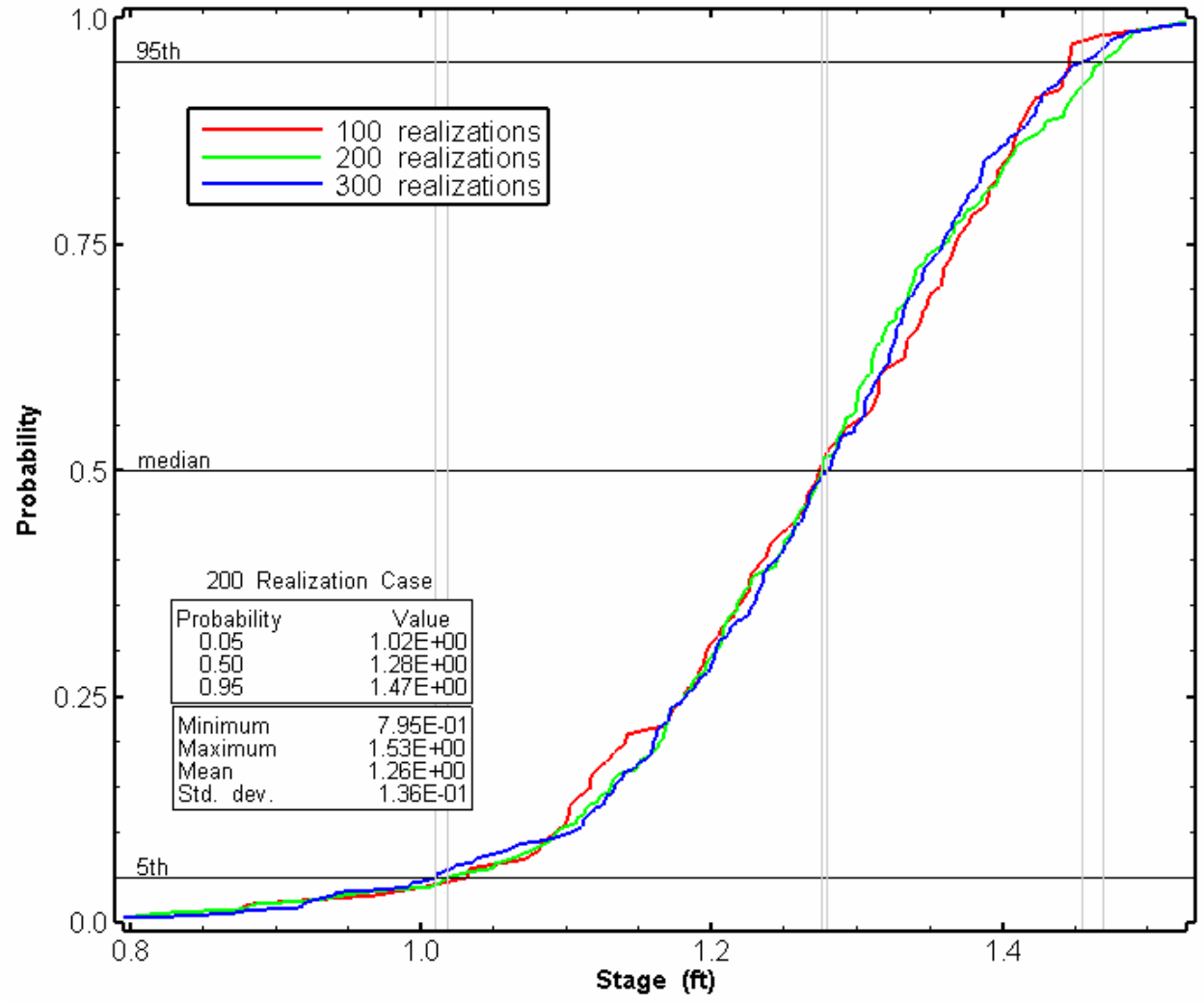


Figure 5-1 Stability analysis for the [25492stage] metric.

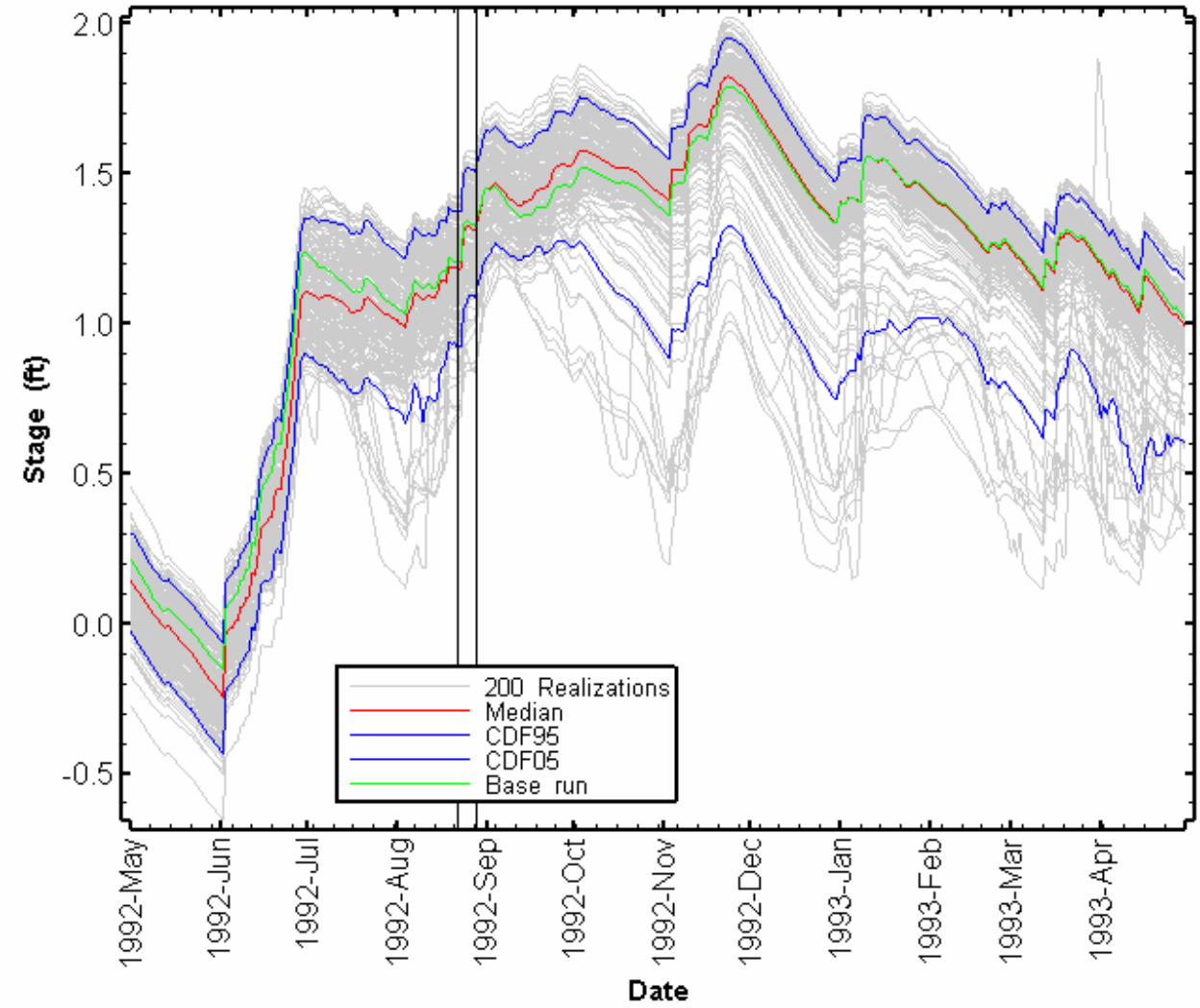


Figure 5-2 Horsetail plot of stage at cell 25492.

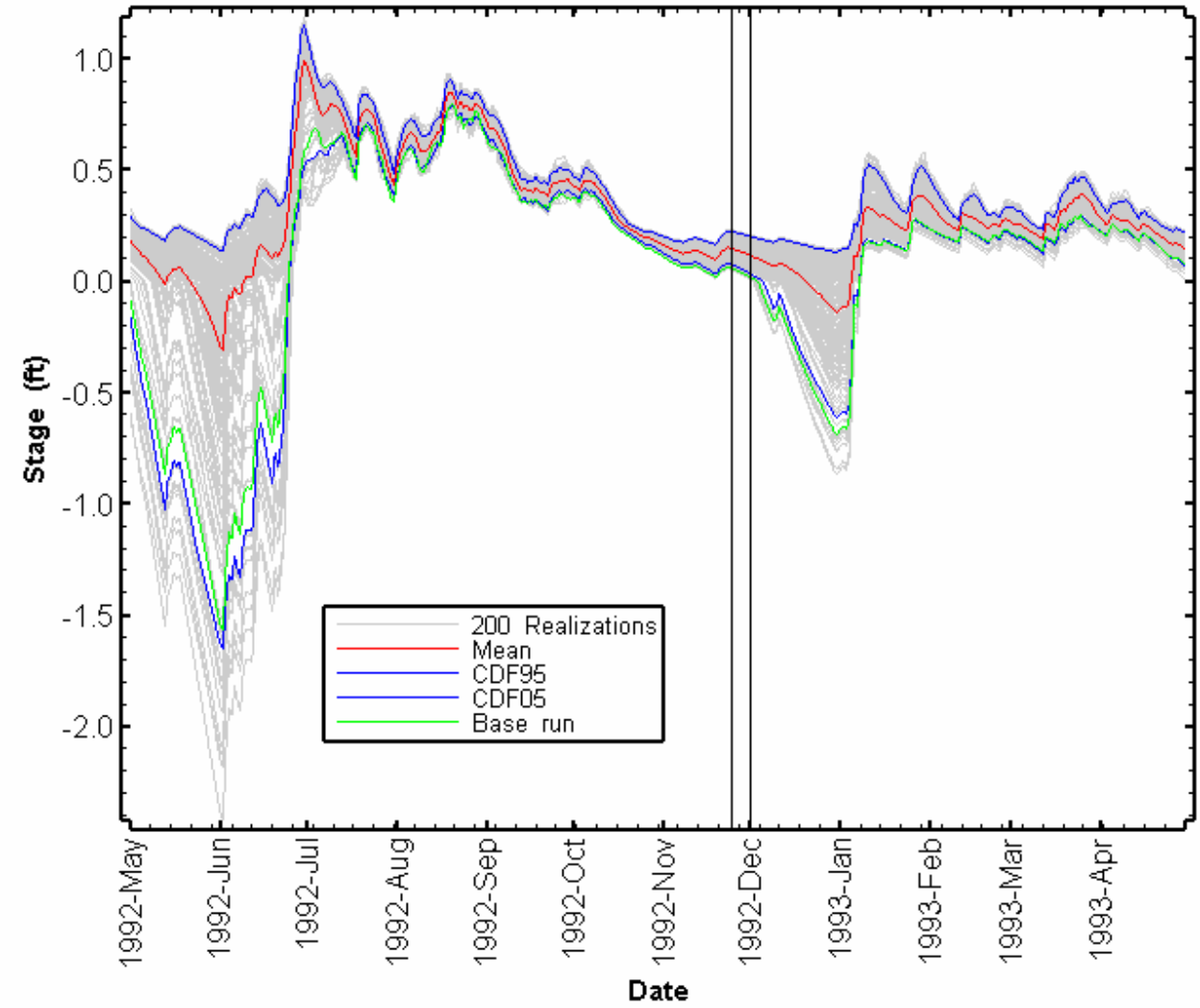


Figure 5-3 Horsetail plot of stage at cell 25087.

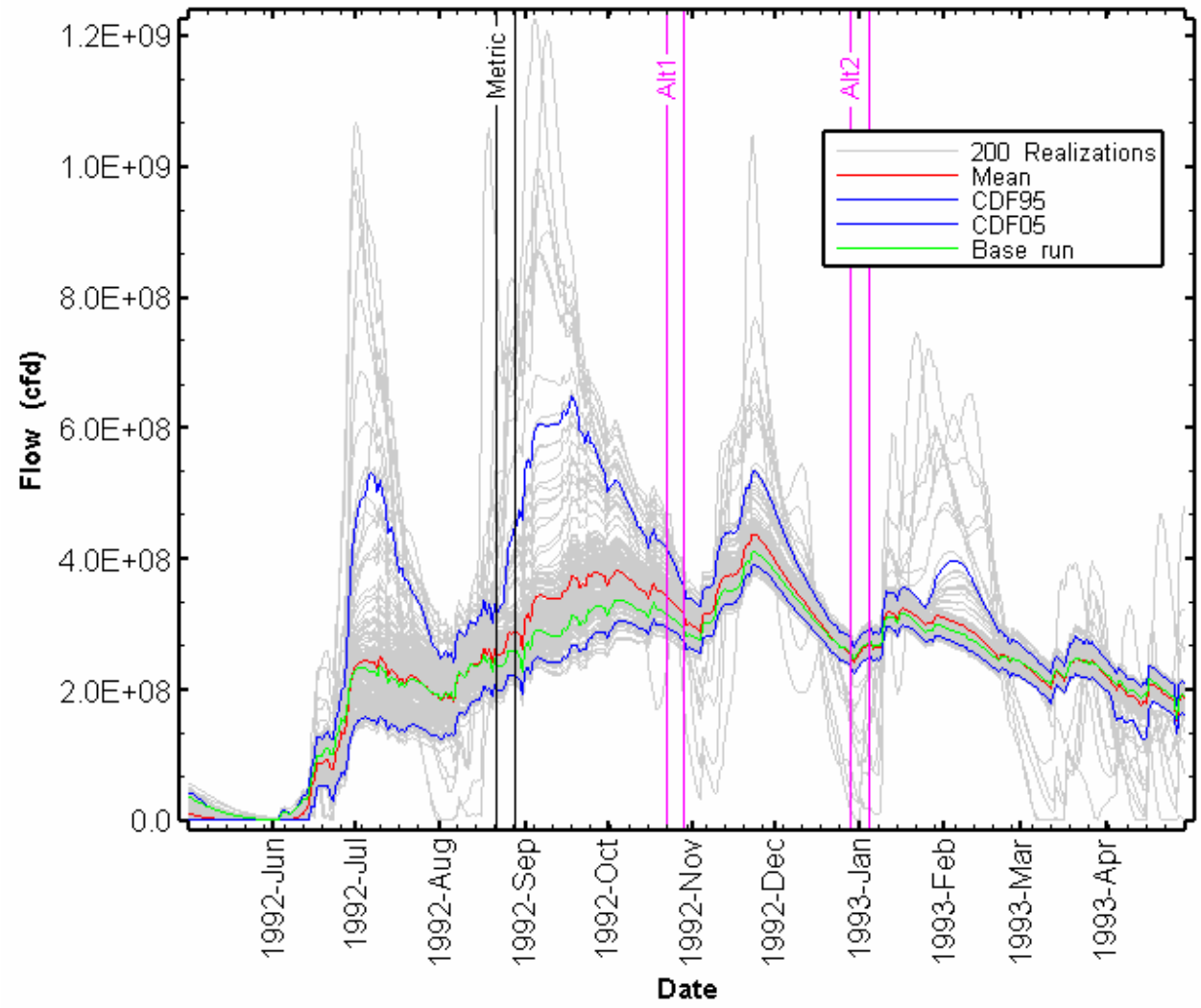


Figure 5-4 Horsetail plot of daily transect flow for Tamiami.

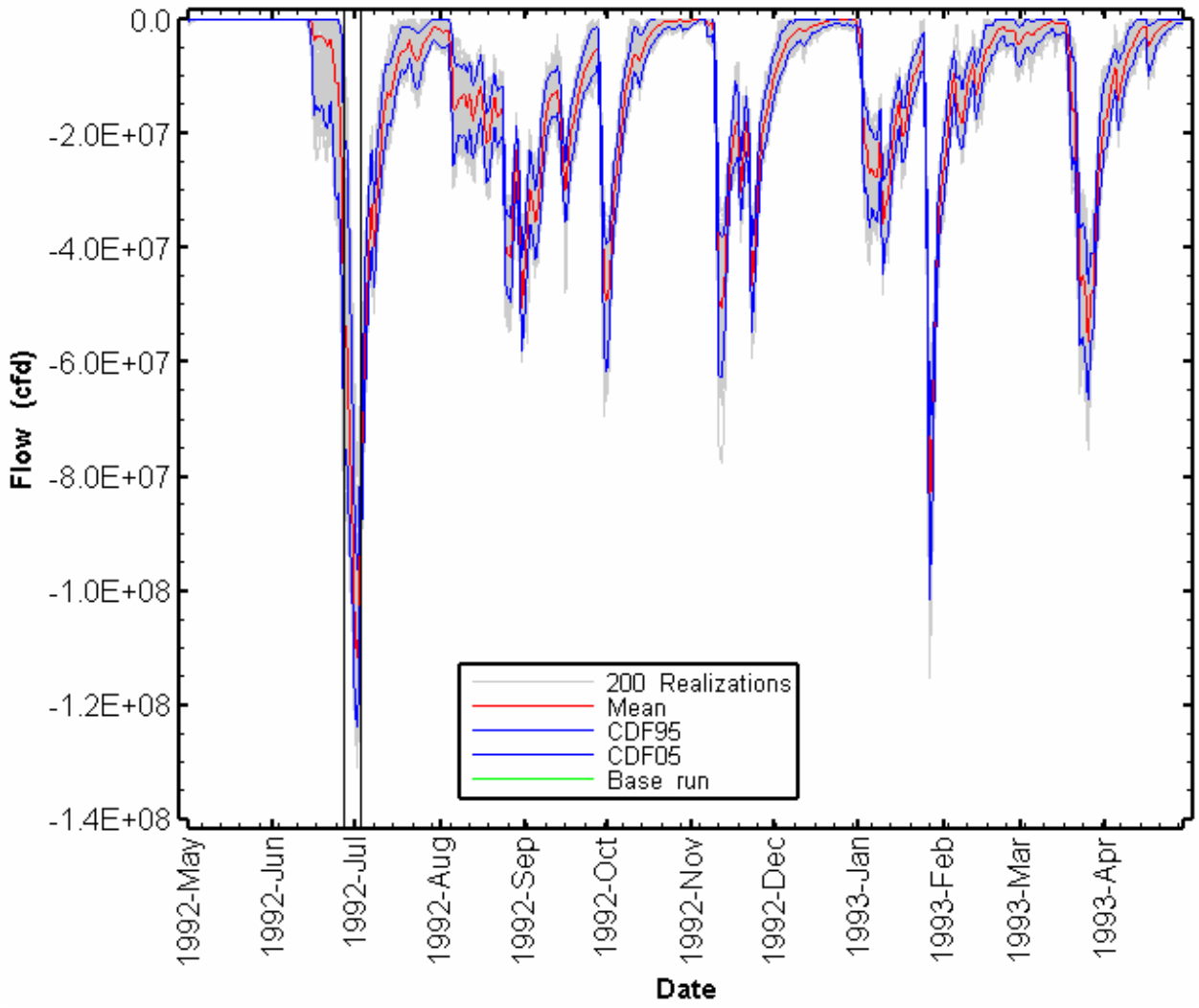


Figure 5-5 Horsetail plot of daily transect flow for T712_East.

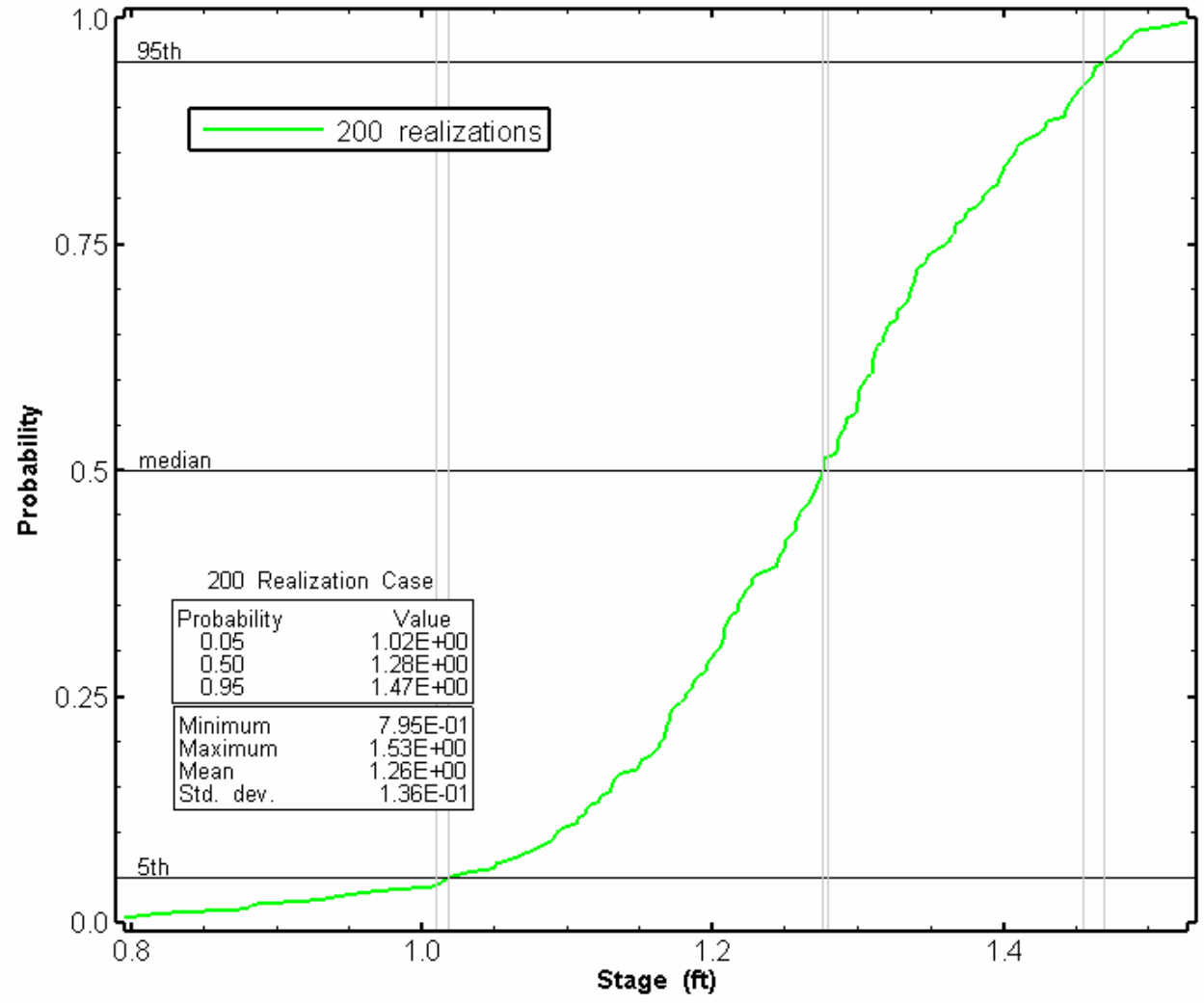


Figure 5-6 CDF for the [25492stage] metric, for the 200 realization case.

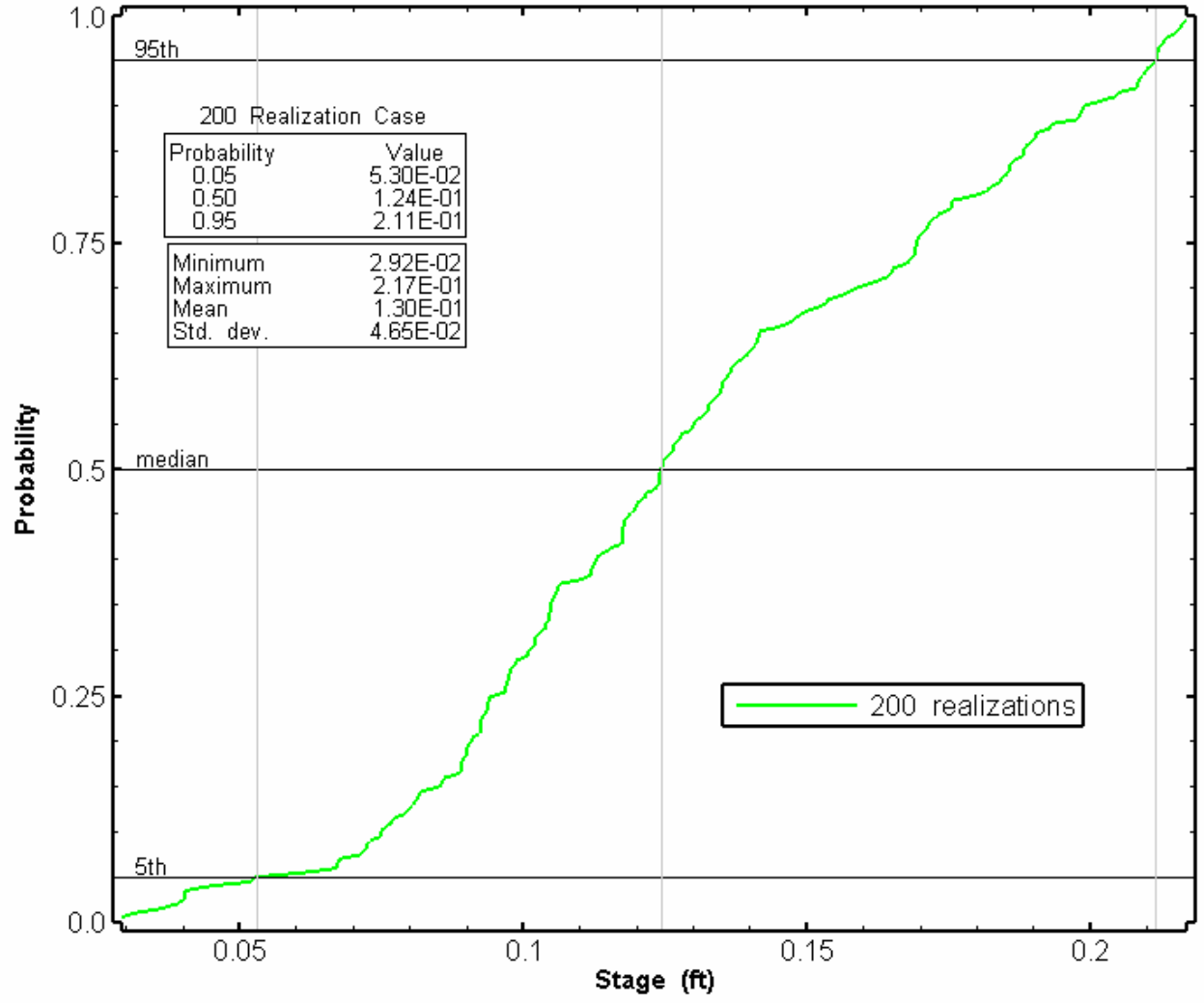


Figure 5-7 CDF for the [25087stage] metric, for the 200 realization case.

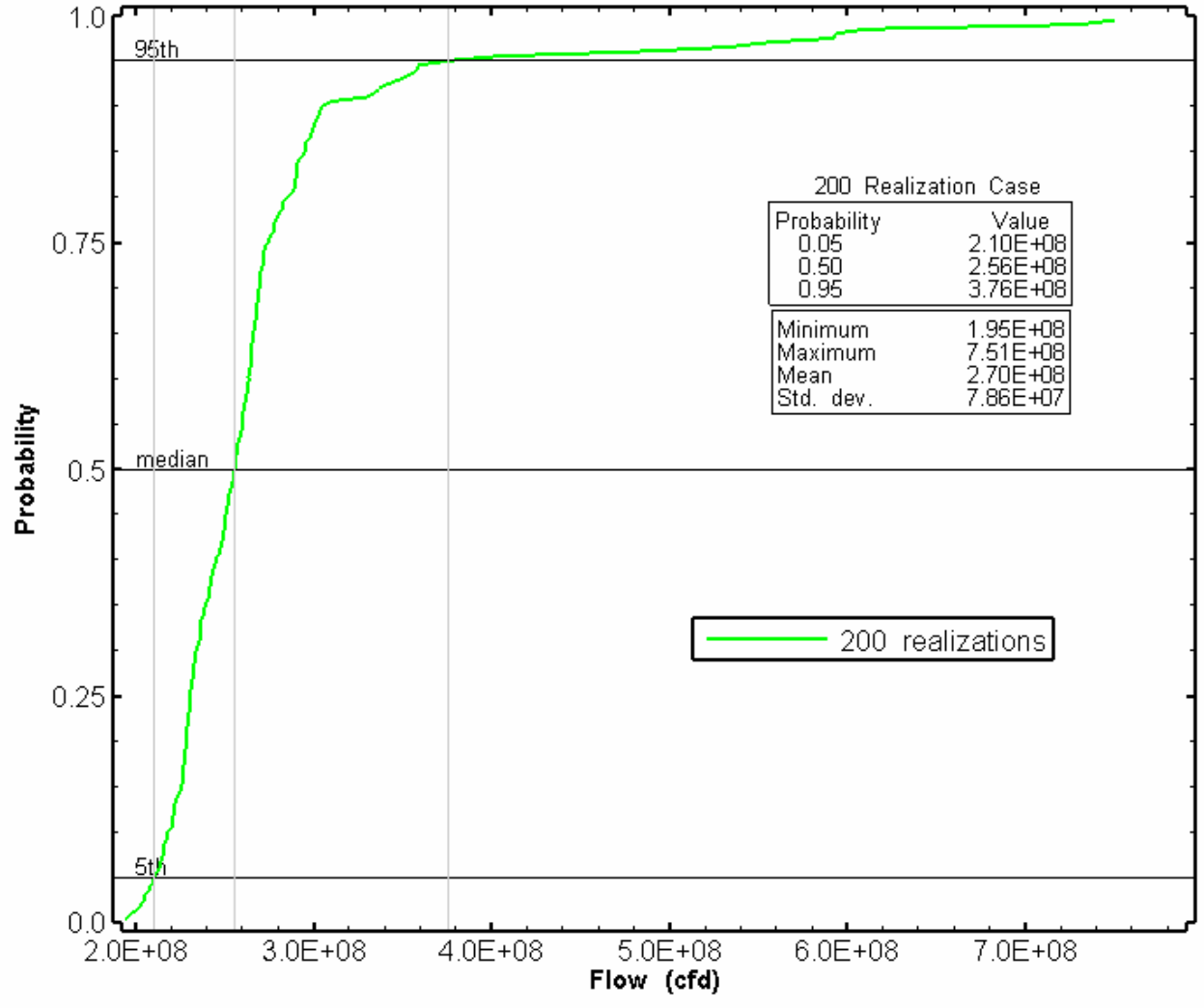


Figure 5-8 CDF for the [Tamiami] metric, for the 200 realization case.

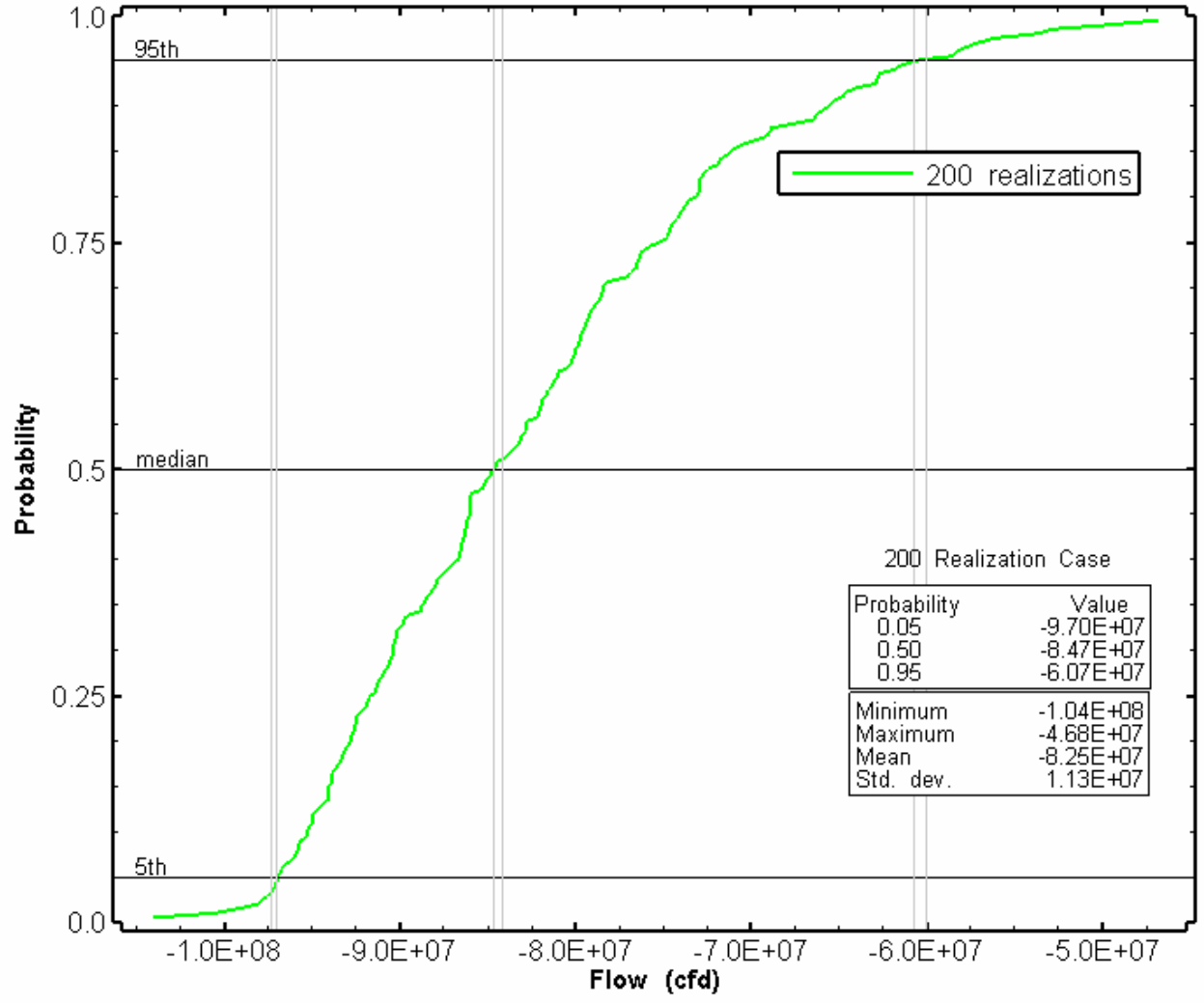


Figure 5-9 CDF for the [T712_East] metric, for the 200 realization case.

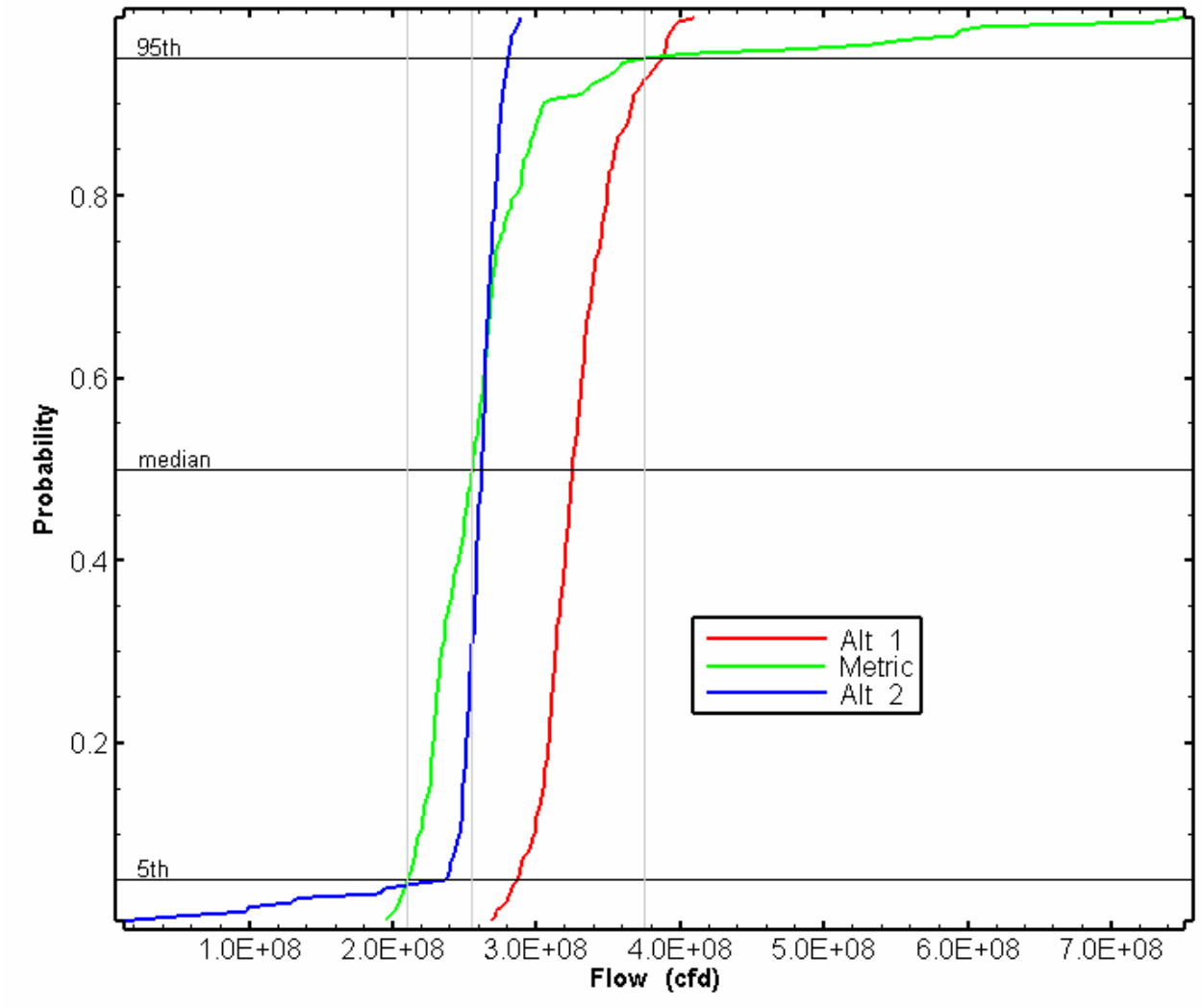


Figure 5-10 CDFs for the [Tamiami] metric, and two alternate metrics, for the 200 realization case. Figure 5-4 shows the time slices for the three metrics.

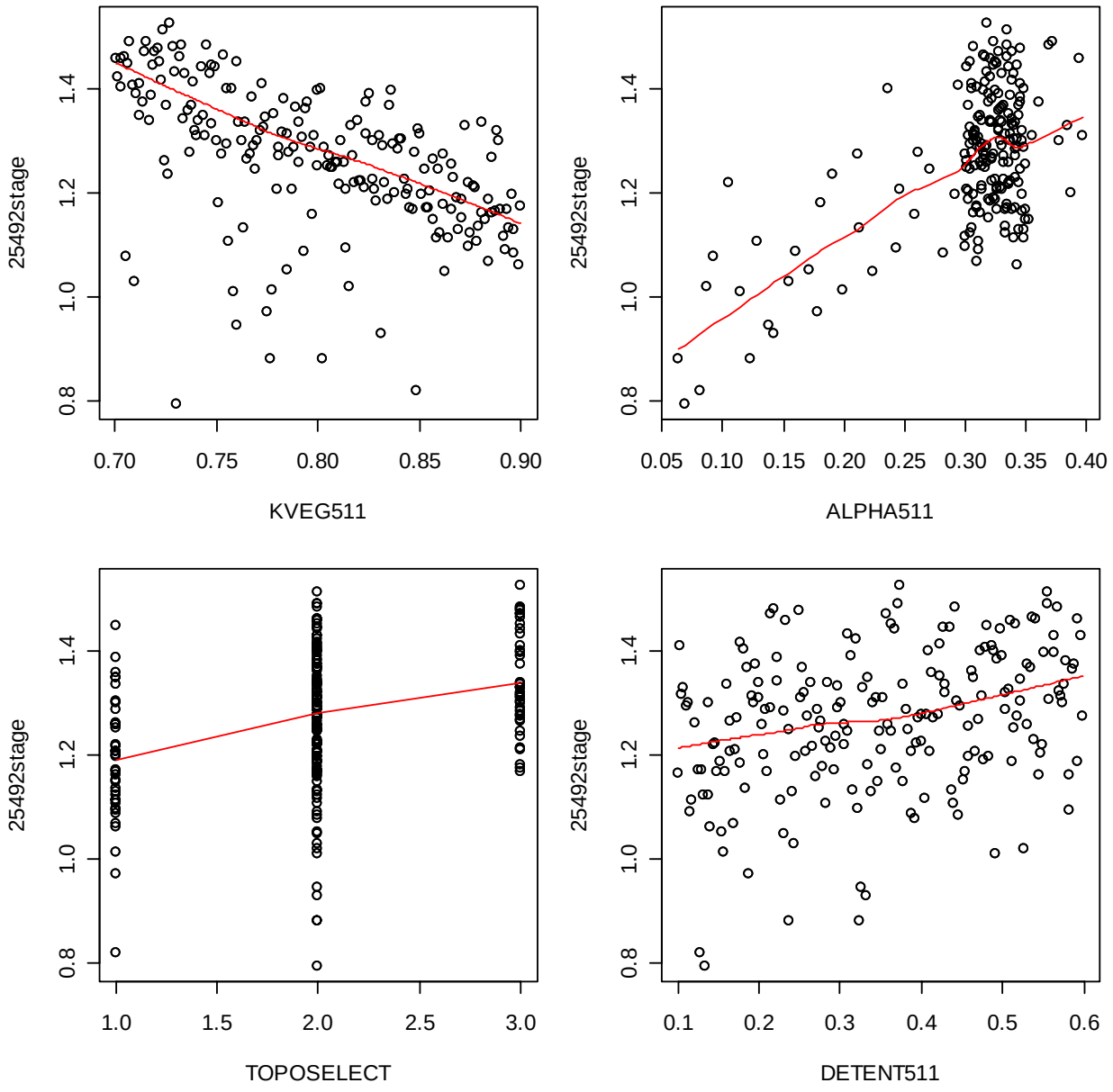


Figure 6-1 Input-output scatterplots for important variables with respect to [25492stage].

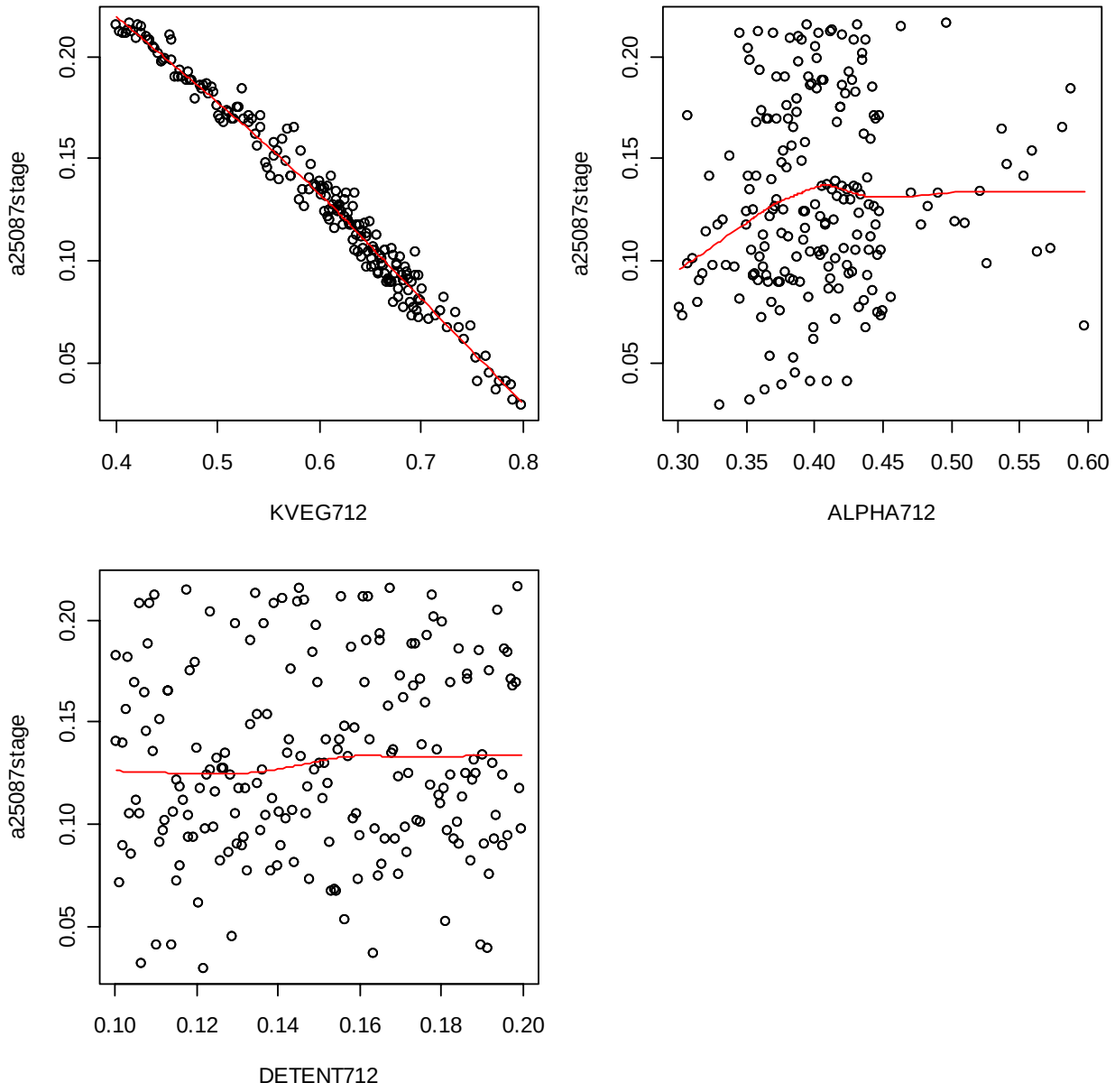


Figure 6-2 Input-output scatterplots for important variables with respect to [25087stage].

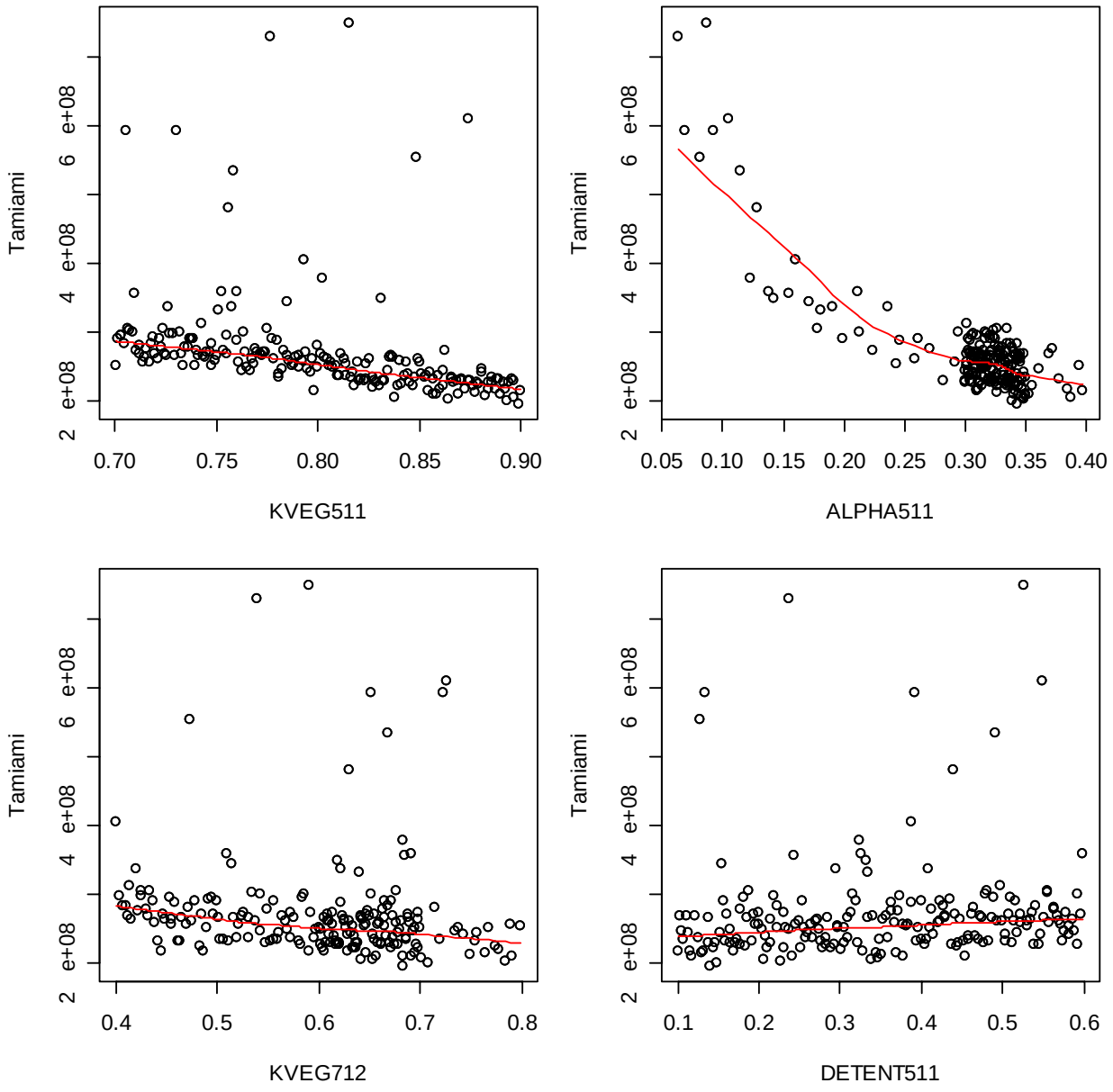


Figure 6-3 Input-output scatterplots for important variables with respect to [Tamiami].

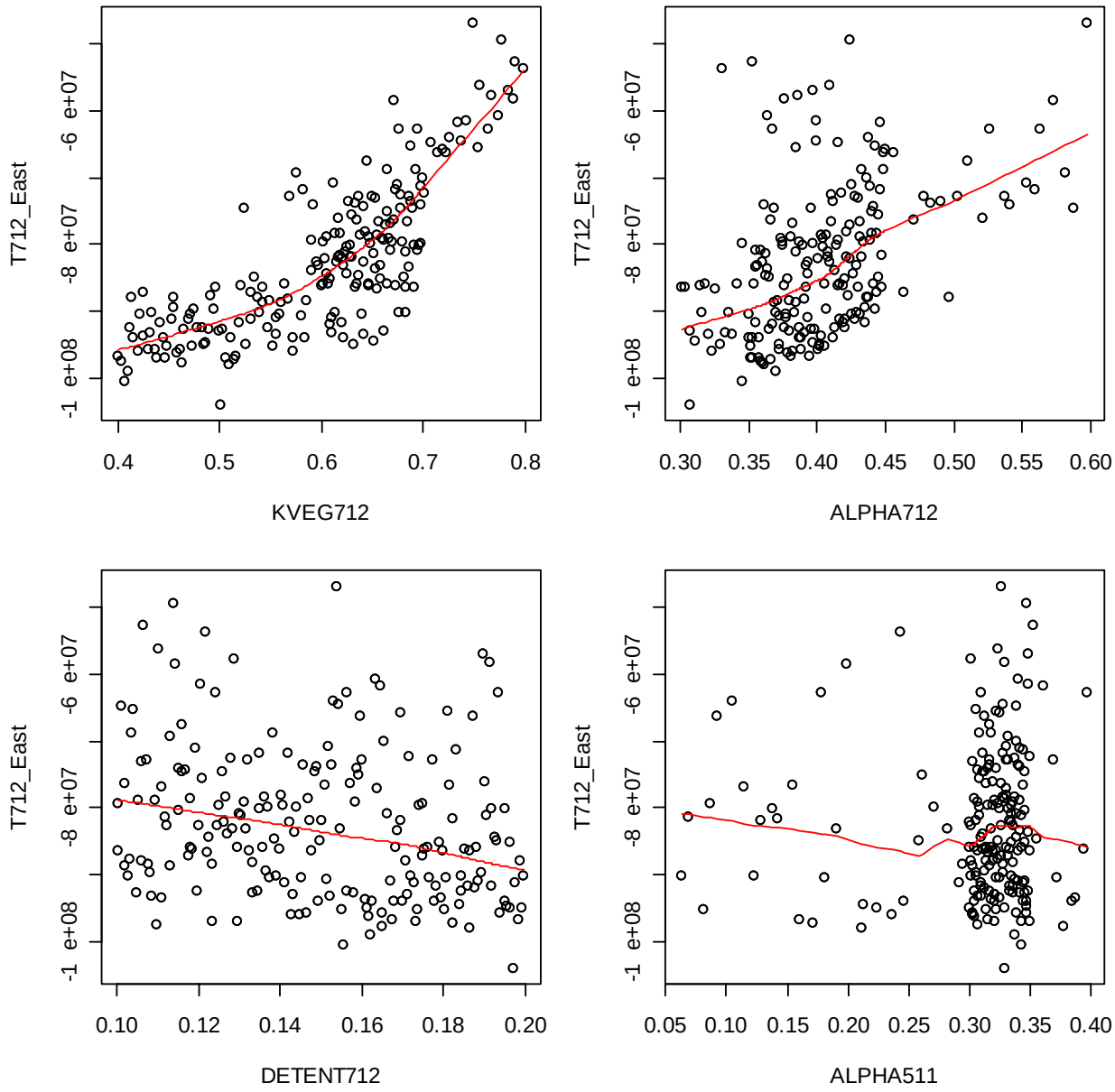


Figure 6-4 Input-output scatterplots for important variables with respect to [T712_East].

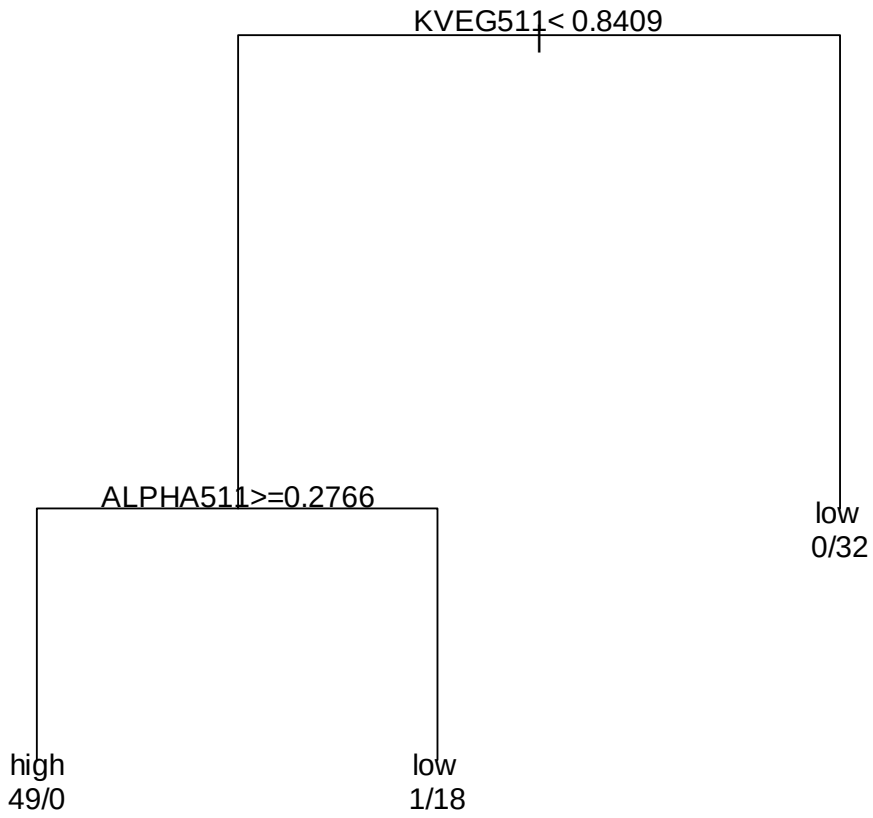


Figure 6-5 Classification tree for metric [25492stage].

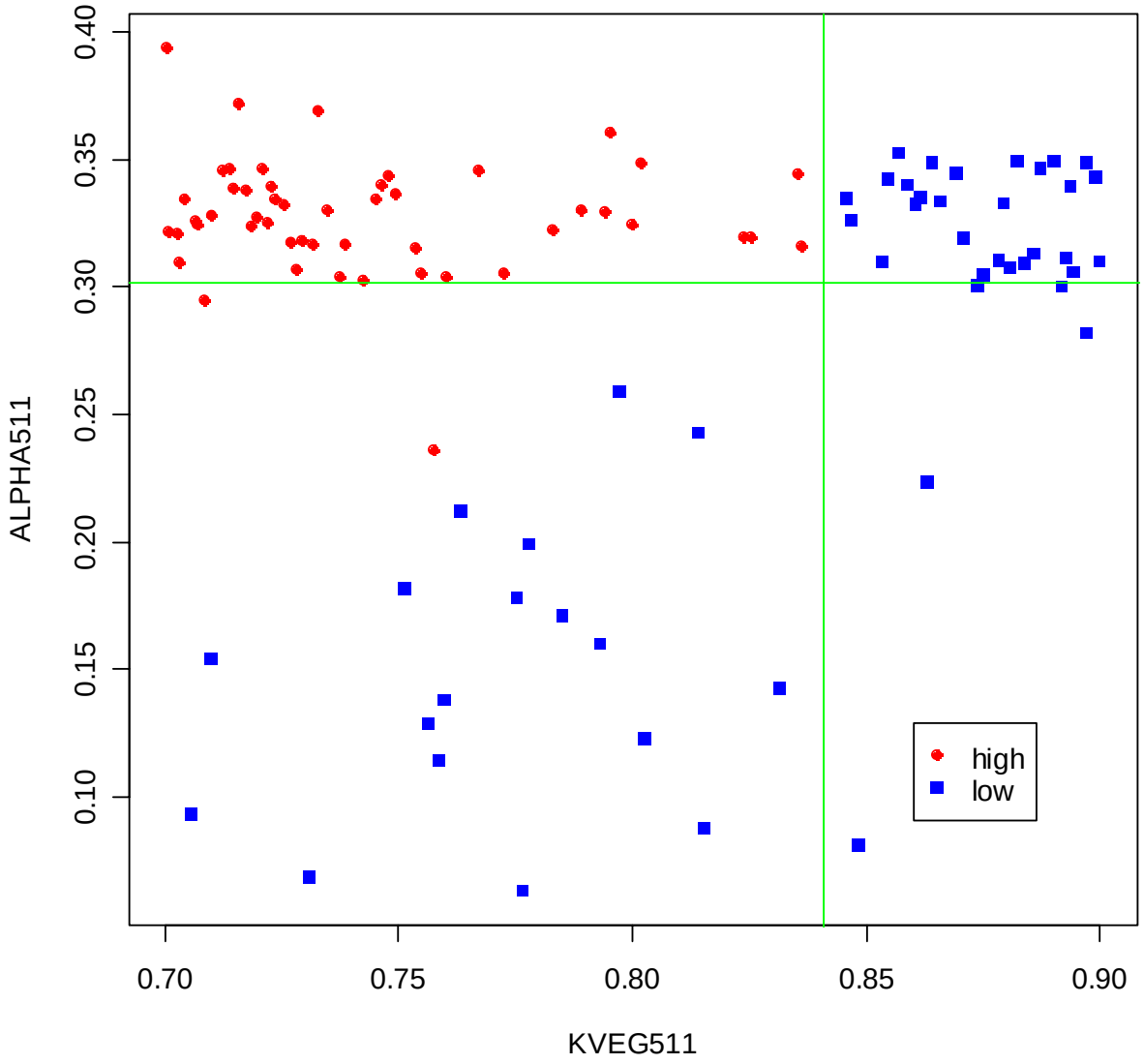


Figure 6-6 Partition plot for metric [25492stage]

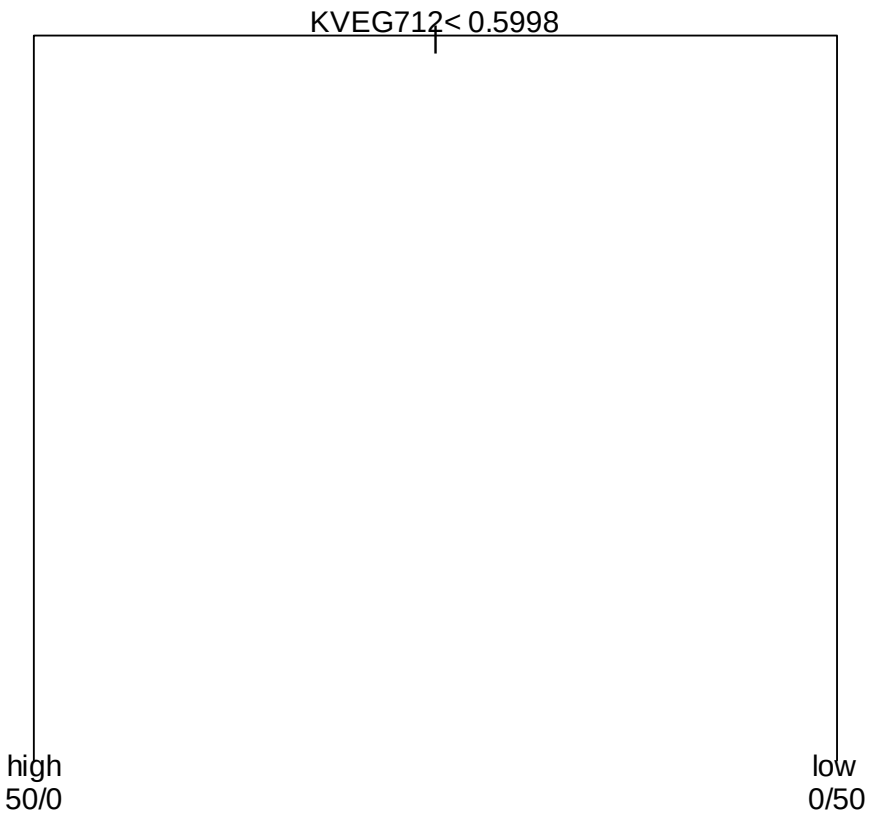


Figure 6-7 Classification tree for metric [25087stage].

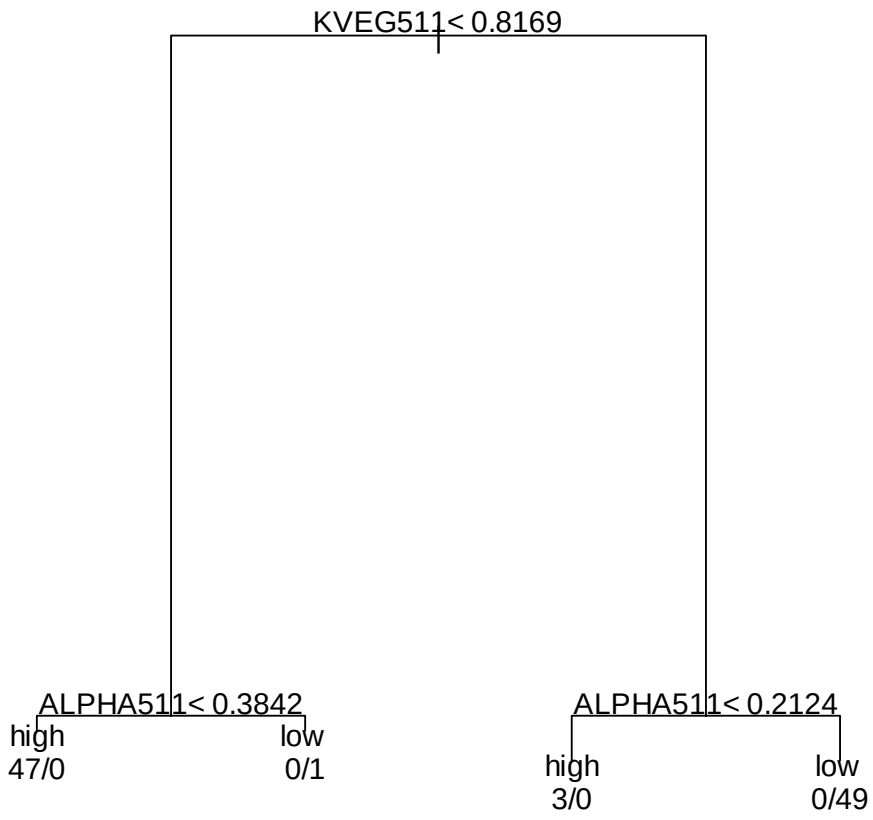


Figure 6-8 Classification tree for metric [Tamiami].

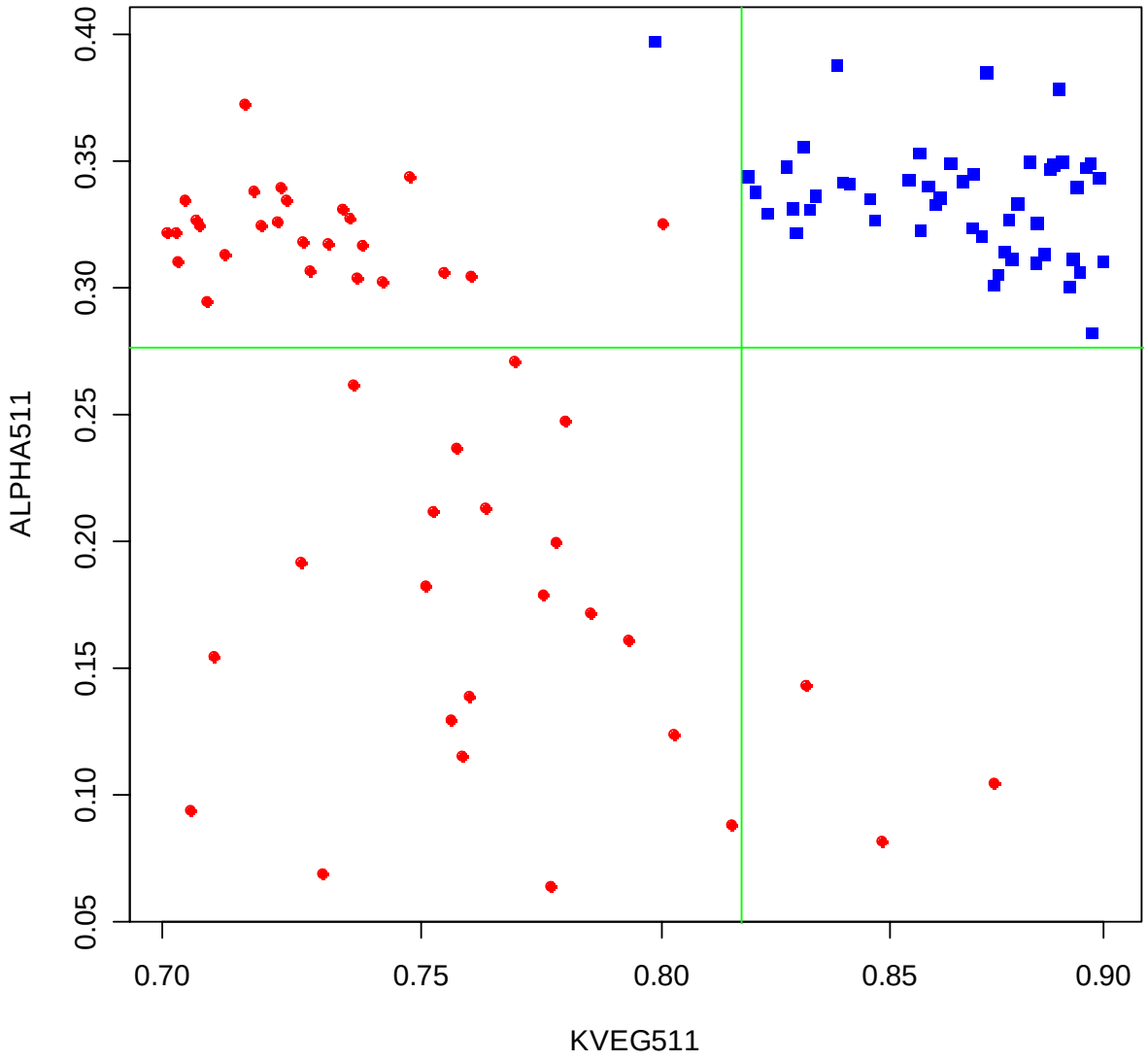


Figure 6-9 Partition plot for metric [Tamiami]

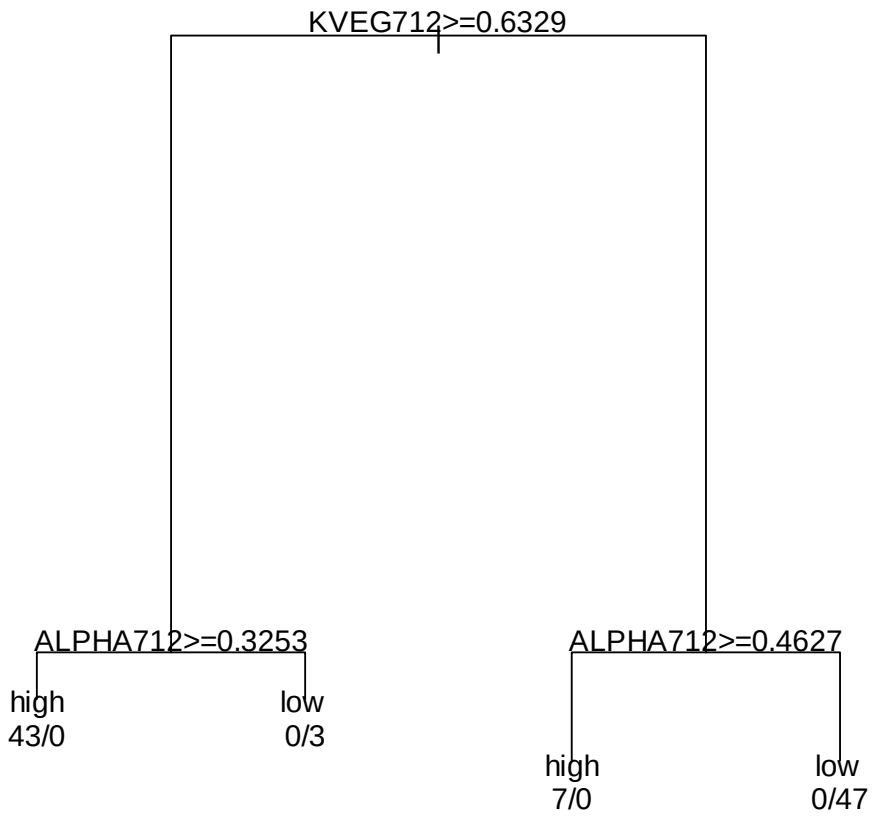


Figure 6-10 Classification tree for metric [T712_East].

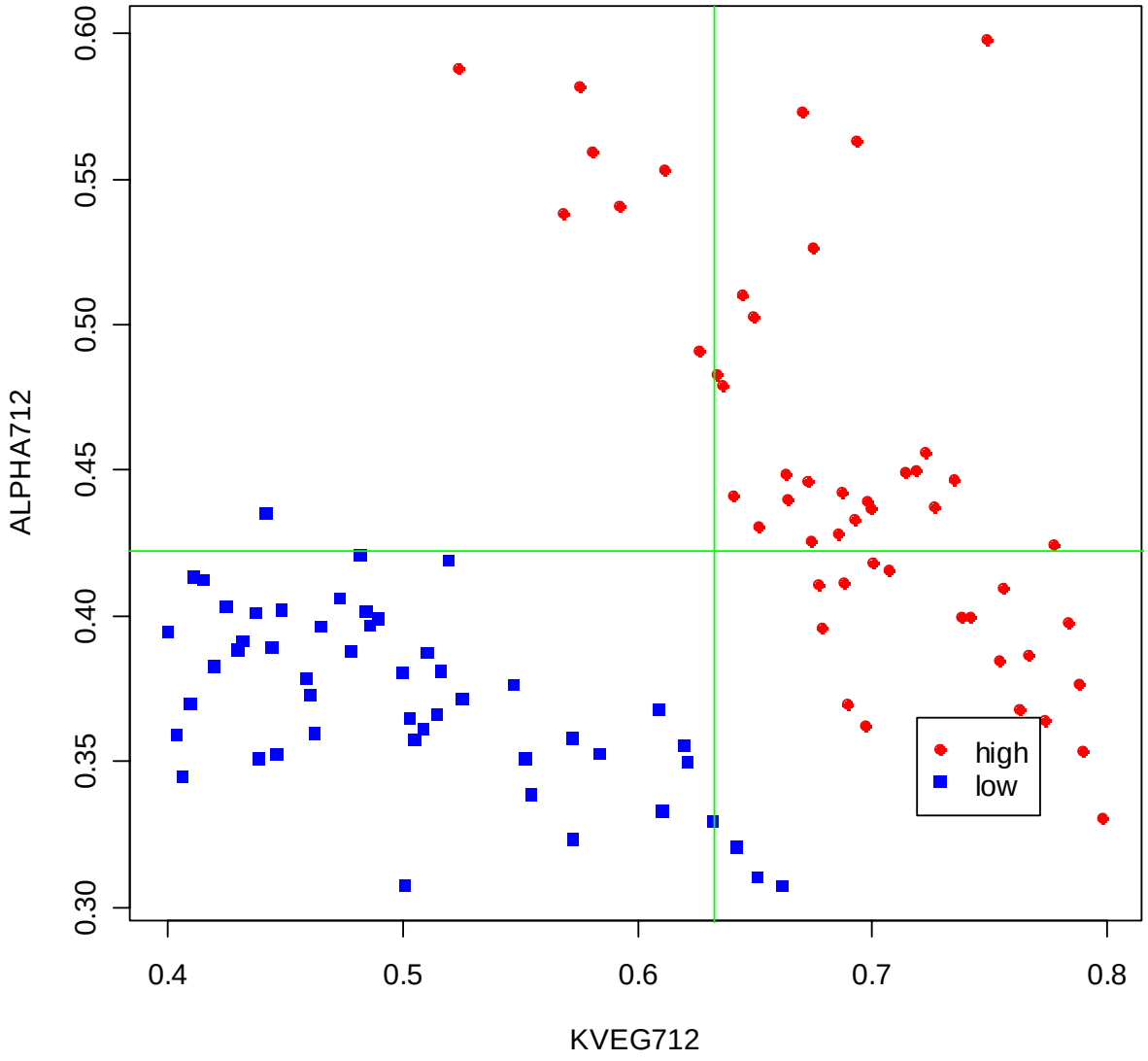


Figure 6-11 Partition plot for metric [T712_East].